A MONTE CARLO STUDY OF MULTIPLE COMPARISON TECHNIQUES

by

Merlin Gene Bell

United States
Naval Postgraduate School

THESIS

A MONTE CARLO STUDY OF
MULTIPLE COMPARISON TECHNIQUES

by

Merlin Gene Bell

October 1969

A Monte Carlo Study of Multiple Comparison Techniques

by

Merlin Gene Bell
Lieutenant, United States Navy
B.S.A.E., Purdue University, 1962

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
October 1969

ABSTRACT

A study was conducted on the multiple comparison methods presented by Scheffé, Tukey, Student-Newman-Keuls, and Duncan under the experimental situation in which all populations were normal with equal variances and all means but one were equal. The characteristics of all four test procedures were compared for the case of multiple comparisons of pairs of means. These tests were conducted both with and without the prior performance of an analysis of variance. The Tukey and Scheffé procedures were compared in tests of linear combinations of three means. Estimates were made of the power of the tests and of Type I error rates under both the null and alternate hypotheses. Scheffé's method was found to be too conservative for pairwise comparisons of means, but it was to be preferred over Tukey's method for combinations of more than two means. Duncan's method was the most powerful test of pairwise comparisons, but it maintained little control over one kind of Type I error. The S-N-K procedure showed a good balance between power and control of Type I errors.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF DRAWINGS

18. Estimated (conditioned) power of 5% Tukey test of four means

19. Estimated (conditioned) power of 5% Tukey test of five means

20. Estimated power of 5% Tukey test of a linear combination of 3 means

21. Estimated (unconditioned) power of 5% S-N-K test of three means

22. Estimated (unconditioned) power of 5% S-N-K test of four means

23. Estimated (unconditioned) power of 5% S-N-K test of five means

24. Estimated (unconditioned) power of 1% S-N-K test of three means

25. Estimated (unconditioned) power of 1% S-N-K test of four means

26. Estimated (unconditioned) power of 1% S-N-K test of five means

27. Estimated (conditioned) power of 5% S-N-K test of three means

28. Estimated (conditioned) power of 5% S-N-K test of four means

29. Estimated (conditioned) power of 5% S-N-K test of five means

30. Estimated (unconditioned) power of 5% Duncan test of three means

31. Estimated (unconditioned) power of 5% Duncan test of four means

32. Estimated (unconditioned) power of 5% Duncan test of five means

33. Estimated (unconditioned) power of 1% Duncan test of three means

34. Estimated (unconditioned) power of 1% Duncan test of four means

35. Estimated (unconditioned) power of 1% Duncan test of five means

I. INTRODUCTION

Consider the experiment in which random samples are drawn from several different populations in order to test for equality of the population means. It is often assumed that these populations are normal and that the population variances are equal but unknown. To be more specific, suppose there are k groups of independent observations $X_{11}, X_{12}, \cdots X_{1n}, X_{21}, X_{22}, \cdots, X_{2n}, \cdots, X_{k1}, X_{k2} \cdots, X_{kn}$, from normally distributed populations with means $\mu_1, \mu_2, \ldots, \mu_k$ and common variance σ^2 (unknown), where X_{ij} represents the outcome of the $j^{\underline{th}}$ sample on the $i^{\underline{th}}$ population. Note that it has been assumed that the population samples are all of size n.

To test the null hypothesis, H_o: $\mu_1 = \mu_2 = \ldots = \mu_k$, the model I analysis of variance is usually employed. The resulting test statistic, formed to test H_o, is:

$$F_{k-1,k(n-1)} = \frac{\text{mean square between groups}}{\text{mean square within groups}}$$

$$= \frac{k(n-1)(n) \sum_{i=1}^{k} (\bar{X}_i - \bar{X})^2}{(k-1) \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2}$$

The null hypothesis is rejected if the test statistic exceeds the critical value appropriate for the significance level of the test.

Suppose the experimenter has rejected the null hypothesis. In many cases he will want to know which means differ

and which means do not. The multiple comparison procedures were designed to help answer just this question. The research, the results of which are presented in this paper, attempted to assess the relative merits of four of the most commonly used techniques. The characteristics of the procedures presented by Scheffé [1953], Tukey [1949], Duncan [1955] and one credited variously to Student, Newman [1939], and Keuls [1952] were studied for the case in which all population means but one were equal. The tests were used to determine if $\mu_i = \mu_j$, $i \neq j$. One brief experiment was conducted to compare the performance of the Scheffé and Tukey methods when testing a hypothesis concerning a linear combination of more than two means.

II. MULTIPLE COMPARISON PROCEDURES

Miller [1966] presents a detailed discussion of the test procedures and their underlying theoretical bases. This work also contains a most complete bibliography of the field of multiple comparisons. The mechanics of the test procedures are presented in the following paragraphs.

$$\text{Let } S^2 = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2 .$$

Then S^2 is an estimator for σ^2 with $k(n-1)$ degrees of freedom. This estimator was used in conjunction with all of the methods throughout the experiment.

A. SCHEFFÉ'S METHOD

Scheffé's method is more general than the other methods studied in that it does not require equal sample sizes from all the groups. It may also be used to test the hypothesis that a general linear function or contrast is zero. The presentation here is for the special case of equal sample sizes from all populations.

Consider the linear function or contrast of the population means,

$$\lambda = \sum_{i=1}^{k} c_i \mu_i$$

where

$$\sum_{i=1}^{k} c_i = 0 .$$

An estimate of the contrast is:

$$L = \sum_{i=1}^{k} c_i \bar{X}_i .$$

Since the $\bar{X}_i$ are independent, the variance of L is given by:

$$\sigma_L^2 = \frac{\sigma^2}{n} \sum_{i=1}^{k} c_i^2 .$$

and its estimate is:

$$S_L^2 = \frac{S^2}{n} \sum_{i=1}^{k} c_i^2 .$$

Scheffé [1953] derived a confidence interval for all possible contrasts, λ, which in turn implies a test of significance for the null hypothesis H_o: $\lambda = 0$. Reject H_o if

$$|L| > \{(k-1)\ F_\alpha\ [k-1, k(n-1)]\ (S_L^2)\}^{\frac{1}{2}}$$

where $F_\alpha[k-1, k(n-1)]$ is the tabulated F-value at the α significance level for $\upsilon_1 = k-1$ and $\upsilon_2 = k(n-1)$.

For the special case where $c_i = -1$, $c_j = 1$, and $c_m = 0$, $m \neq i$ or j, $\lambda = \mu_j - \mu_i$. The test criterion for H_o: $\mu_i = \mu_j$, $i, j = 1, 2, \ldots, k$, $i \neq j$ then becomes: reject H_o if

$$\bar{X}_j - \bar{X}_i > \{(k-1)\ F_\alpha\ [k-1,\ k(n-1)]\ (S_L^2)\}^{\frac{1}{2}},$$

where the population sample means are ranked in ascending order and $\bar{X}_j > \bar{X}_i$. This may be rewritten as:

$$\bar{X}_j - \bar{X}_i > \{(k-1)\ F_\alpha\ [k-1, k(n-1)]\ (\frac{2S^2}{n})\}^{\frac{1}{2}}.$$

B. TUKEY'S METHOD

Tukey's method, in contrast to that of Scheffé, was designed primarily for tests of simple differences of means, $\mu_j - \mu_i$, although it too is sufficiently general to be used for tests of linear combinations of means. It is exact only for equal sample sizes from all groups; however, modifications have been proposed to allow its use in the case of unequal sample sizes [Bancroft, 1968]. As above, let

$$\lambda = \sum_{i=1}^{k} c_i \mu_i$$

be estimated by

$$L = \sum_{i=1}^{k} c_i \bar{X}_i$$

where

$$\sum_{i=1}^{k} c_i = 0 .$$

For this test the pivotal test statistic is the studentized range rather than the F statistic of the previous method. Tukey [1949] has shown that the test of H_o: $\lambda = 0$ is performed by rejecting H_o if

$$|L| > Q_\alpha [k, k(n-1)] \left(\frac{S^2}{n}\right)^{\frac{1}{2}} \sum_{i=1}^{k} \frac{|c_i|}{2} ,$$

where $Q_\alpha[k, k(n-1)]$ is the upper 100α percent point of the studentized range distribution with parameters k and k(n-1).

For the special case of pairwise mean comparisons, with the population means placed in ascending order and $\bar{X}_j > \bar{X}_i$,

$$|L| = \bar{X}_j - \bar{X}_i ,$$

and

$$\sum_{i=1}^{k} \frac{|c_i|}{2} = 1 .$$

The test criterion for this case is then

$$\bar{X}_j - \bar{X}_i > Q_\alpha [k, k(n-1)] \left(\frac{S^2}{n}\right)^{\frac{1}{2}} .$$

C. STUDENT-NEWMAN-KEULS (S-N-K) PROCEDURE

The last two tests to be presented are both classified as multiple range procedures. These procedures are not adaptable for tests of general linear combinations of the population means. The S-N-K procedure was first proposed by Newman [1939] and independently by Keuls [1952]. The

basic idea has been attributed to Student (W. S. Gosset) [Miller, 1966].

The null hypothesis H_o: $\mu_1 = \mu_2 = \ldots = \mu_k$ is to be tested against the alternative H_1: $\mu_i \neq \mu_j$, $i \neq j$, $i, j = 1, 2, \ldots, k$. The procedures were designed to declare which means are significantly different if H_o is rejected. The S-N-K test procedure is as follows:

1. Arrange the sample means in ascending order of magnitude as $\bar{X}_{(1)}, \bar{X}_{(2)}, \ldots, \bar{X}_{(k)}$.
2. Calculate the p-mean critical differences for $p = 2, 3, \ldots, k$,

$$C_p = Q_\alpha[p, k(n-1)]\left(\frac{S^2}{n}\right)^{\frac{1}{2}},$$

where $Q_\alpha[p, k(n-1)]$ is the upper 100α percent point of the studentized range distribution with parameters p and $k(n-1)$.

3. Declare $\mu_{(1)}$ and $\mu_{(k)}$ significantly different if $\bar{X}_{(k)} - \bar{X}_{(1)} > C_k$. If $\mu_{(1)}$ and $\mu_{(k)}$ do not differ significantly accept H_o and stop the procedure.
4. Declare $\mu_{(k)}$ different from $\mu_{(2)}$ if $\bar{X}_{(k)} - \bar{X}_{(1)} > C_k$ and $\bar{X}_{(k)} - \bar{X}_{(2)} > C_{k-1}$. If this does not hold then state that $\mu_{(k)}$ does not differ significantly from $\mu_{(2)}, \mu_{(3)}, \ldots, \mu_{(k-1)}$, and hence any pair of means μ_i, μ_j, $i, j = 2, 3, \ldots, k$ is not significantly different. Similarly declare $\mu_{(k-1)}$ different from $\mu_{(1)}$ if $\bar{X}_{(k)} - \bar{X}_{(1)} > C_k$ and $\bar{X}_{(k-1)} - \bar{X}_{(1)} > C_{k-1}$; otherwise, state that $\mu_{(k-1)}$ does not differ significantly from $\mu_{(1)}, \mu_{(2)}, \ldots, \mu_{(k-2)}$, and hence any pair of means μ_i, μ_j, $i, j = 1, 2, \ldots, k-1$ is not significantly different.

5. Proceed, if necessary, until all groups of size p, p = 2, 3, ..., k have been declared. Note that once an ordered subset P of means of size p has been declared not significant, all ordered subsets of P of sizes p-1, p-2, ..., 2 also must be declared nonsignificant.

D. DUNCAN PROCEDURE

Duncan [1955] proposed a modification of the S-N-K multiple range procedure with less conservative critical values. The test procedure is the same as that outlined above for the S-N-K method except that in step (2) the p-mean critical differences are defined by

$$C_p = Q_{\alpha_p}[p, k(n-1)]\left(\frac{s^2}{n}\right)^{\frac{1}{2}},$$

where $Q_{\alpha_p}[p, k(n-1)]$ is the upper $100\alpha_p$ percent point of the studentized range with parameters p and k(n-1) and $\alpha_p = 1 - (1-\alpha)^{p-1}$. These special percentage points of the studentized range have been tabulated by Harter, Clemm, and Guthrie [1959]. Approximations exist for both the S-N-K and Duncan procedures for the case of unequal group sizes [Sarhan and Greenberg, 1962].

III. ERROR BASES

In contrast to the situation in which a statistical test is performed on two population means, there is no universally accepted measure of the relative effectiveness of two statistical tests such as that provided by the Neyman-Pearson theory. Most if not all the definitions offered in the literature are but intuitive extensions of the Neyman-Pearson ideas. There are three possible types of errors, and in this paper the following definitions were used:

1. A Type I error was said to have occurred if μ_i was declared different from μ_j when, in fact, $\mu_i = \mu_j$.
2. A Type II error was said to have occurred if μ_i was declared not different from μ_j when, in fact, $\mu_i \neq \mu_j$.
3. A Type III error was said to have occurred if μ_i was declared greater than μ_j when, in fact, $\mu_i < \mu_j$. Type III errors were not tabulated in the experiments and the definition is given only for the sake of completeness.

A. TYPE I ERRORS

Following the definitions of Bancroft [1968], two definitions of Type I errors were used in this paper. The per-comparison error rate was defined as the long-run value of

$$\frac{\text{Number of comparisons falsely declared significant}}{\text{Total no. of comparisons in which no true difference existed}},$$

and the experimentwise error rate was defined as the long-run value of

$$\frac{\text{Number of experiments in which at least one difference was falsely declared significant}}{\text{Total number of experiments}}.$$

When testing for equality of means among more than two population means, Type I errors can occur when H_o is true and also when H_o is false.

B. TYPE II ERRORS

In this study Type II errors were not assessed directly, but instead the concept of power was used. The power of a test for a specified configuration of the population means was defined as the long-run value of

$$\frac{\text{Number of comparisons correctly declared significant}}{\text{Number of comparisons in which a true difference existed}}.$$

IV. THE EXPERIMENTS

All of the experiments were simulated on an IBM 360/67 computer. Standard normal variates were generated using the Naval Postgraduate School computer facility library Gaussian Random Number Generator (GPN) which is based on the general scheme devised by Marsaglia [1964]. The uniform random numbers required in the routine were generated using an additive-congruential method tested by Green, Smith, and Klem [1959]. This method started with sixteen random numbers, $X_1, \ldots, X_{16}$ and generated the sequence of random numbers $X_j = (X_{j-1} + X_{j-16}) \bmod 1$, $j > 16$. The Gaussian Random Number Generator has been tested for accuracy by taking means, deviations, skewness, and kurtosis on 35 samples of 10,000 numbers. The results showed that the routine generated distributions with normal characteristics.[1]

The critical values used in all of the experiments were obtained by linear harmonic interpolation (where necessary) of the tabulated values. Percentage points of the (Snedecor) F distribution were obtained from those calculated by Merrington and Thompson [1943]. Percentage points of the studentized range and critical values for Duncan's Multiple Range Test were taken from the values derived by Harter, Clemm, and Guthrie [1959].

---

[1]Source: NPS Computer Facility, where complete test results are available on file.

A. UNCONDITIONAL COMPARISONS

The first set of experiments was designed to make possible an evaluation of all four multiple comparison techniques. Only contrasts which were simple differences between means were considered, as these are the only contrasts that the S-N-K and Duncan procedures were designed to test.

A data set, consisting of a sample of size n (n = 6, 8, 10, 20, 30, or 40) from each of k (k = 3, 4, or 5) standard normal populations was generated using the Gaussian Random Number Generator. The sample means and the estimate of the common variance were calculated, and the data set was tested using each of the four methods. For each method a comparison was recorded as incorrect if the test declared μ_i different from μ_j since all means were, in fact, equal; otherwise nothing was recorded. If after all comparisons were made there was at least one incorrect comparison for a method, an experimentwise error was recorded for the method.

After completion of the tests with all population means equal, one of the population means was made greater than the others. This change was effected by simply adding 0.2 to the sample mean from population three, $\bar{X}_3$. The modified set of sample means $\{\bar{X}_1, \bar{X}_2, \bar{X}_3', \ldots, \bar{X}_k\}$, where $\bar{X}_3' = \bar{X}_3 + 0.2$, was again subjected to test by each of the four procedures. For each method a comparison was recorded as incorrect if the test declared μ_i different from μ_j when neither i = 3 nor j = 3. If the test declared μ_i different from μ_j and

either i = 3 or j = 3, the comparison was recorded as correct. Experimentwise errors were recorded as before.

In a similar manner, the mean of population three was increased in steps of size 0.2 to a maximum value of 2.0, and the tests were repeated at each step. Upon completion, this made up one replication of each of eleven experiments (one experiment for each value of μ_3). Note that the only difference between the data tested at the various stages was the value of $\bar{X}_3$. The other sample means remained unchanged.

The experiments were repeated for a total of 500 replications for each value of n and k, and then estimates of power and error rates were calculated. This procedure was repeated for sample sizes of n = 6, 8, 10, 20, 30, and 40, for each value of k = 3, 4, and 5, using both 5% and 1% critical values.

1. Error Rate and Power Calculations

The parameter d was defined as the true difference $\mu_3 - \mu_i$, i = 1, 2, ..., k, i ≠ 3. Power was then tabulated as a function of d. The power and error rate estimates were calculated as follows:

1) For d = 0 (H_o true),

$$\text{Per-comparison error rate} = \frac{\text{No. of "incorrect" comparisons}}{\binom{k}{2} \times 500},$$

and

$$\text{Experimentwise error rate} = \frac{\text{No. of experimentwise errors}}{500}.$$

2) For d ≠ 0 (H_o false),

$$\text{Per-comparison error rate} = \frac{\text{No. of "incorrect" comparisons}}{\binom{k-1}{2} \times 500},$$

$$\text{Experimentwise error rate} = \frac{\text{No. of experimentwise errors}}{500},$$

and

$$\text{Power (d)} = \frac{\text{No. of "correct" comparisons when } \mu_3 = d}{(k-1) \times 500}.$$

B. CONDITIONAL COMPARISONS

In many actual applications the investigator first tests his data for equality of means with the appropriate analysis of variance procedure. Only upon rejection of the hypothesis that all means are equal does he look to a multiple comparison technique for an answer to the question, "How do the means differ?" In an attempt to ascertain how this procedure changes the error rates and power function and whether it is advisable to perform the analysis of variance before proceeding, a second set of experiments was conducted. In these experiments only those sets of data which were declared significant by a model I analysis of variance procedure were tested using the four multiple comparison techniques. In other words, only those samples which the investigator adhering to the philosophy described above would have tested were subjected to test. The test procedure and configurations of the means were the same as those described for the unconditional comparisons. Sufficiently many sets of data were generated to provide for exactly 500 experiments for each configuration of the means,

$\mu_1 = 0$, $\mu_2 = 0$, $\mu_3 = d$, ..., $\mu_k = 0$, $d = 0, 0.2, 0.4, \ldots,$ 2.0, and consequently the calculations were made in the manner described for the unconditional comparisons.

For each value of d, the first 500 sets of data failing the F-test were used for test by the multiple comparison procedures. For small values of d the generated data sets were less likely to fail the F-test than those for large values of d. Consequently the sets of sample means tested by the multiple comparisons procedures for different values of d differed not only in the value of $\bar{X}_3$ but also in the values of the other sample means. Recall that in the unconditional comparisons only the value of $\bar{X}_3$ changed for different values of d (for any fixed values of n and k). This procedure seemed more advisable than using only the 500 sets of data which failed the F-test when $d = 0$ and then incrementing $\bar{X}_3$ in steps of size 0.2 in those sets of means.

C. CONTRASTS OTHER THAN SIMPLE DIFFERENCES OF TWO MEANS

Scheffé [1953] has pointed out that although the Tukey procedure gives shorter confidence intervals than the Scheffé procedure for contrasts which are simple differences of two means, $\underline{C}^T = (0, 0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)$ the opposite situation may hold for other linear combinations of the means. For comparison one set of 500 experiments, all at the 5% significance level, was performed on samples of size n from three normal populations to test contrasts of

the form (-2, 1, 1), (1, -2, 1) and (1, 1, -2). Population i was N (μ_i, 1), where $\mu_1 = \mu_2 = 0$ and $\mu_3 = d$, $d = 0, 0.2, \ldots, 2.0$. The eleven different configurations of the means thus produced sixteen different values for the contrasts,

$$\lambda = \sum_{i=1}^{3} c_i \mu_i = \underline{C}^T \underline{\mu} ,$$

for each value of n = 6, 8, 10, 20, 30, 40.

After generation of the data sets, the three contrasts described above were formed in succession and were subjected to test using the Scheffé and Tukey procedures. The confidence intervals for the contrasts formed by the two different methods were tested for inclusion of zero, and the results were recorded.

1. Error Rate and Power Calculations

Type I errors, defined as declaring $\lambda \neq 0$ when $\lambda = 0$, were possible only when $d = 0$. For $d \neq 0$ all of the contrasts were different from zero. The estimates of Type I error rates for each value of n were calculated as follows:

$$\text{Per-comparison error rate} = \frac{\text{Number of contrasts declared different from zero when } d=0}{1500}$$

$$\text{Experimentwise error rate} = \frac{\text{Number of experiments in which at least one of the contrasts declared } \neq 0 \text{ when } d = 0}{500}$$

The estimate of power was calculated as a function of the absolute value of the true value of the contrast, λ, as follows:

$$\text{Power } (|\lambda|) = \frac{\text{Number of contrasts declared different from zero when true value} = \lambda}{\text{Number of tests of contrasts when true value} = \lambda}$$

The possible non-zero values of λ were -0.2, -0.4, ..., -2.0, -2.4, -2.8, ..., -4.0, and due to the nature of the experimental procedure not all values occurred with the same frequency. The values 0.4, 0.8, 1.2, 1.6, 2.0 occurred 1500 times; 0.2, 0.6, 1.0, 1.4, 1.8 occurred 1000 times; and 2.4, 2.8, 3.2, 3.6, 4.0 occurred 500 times.

V. RESULTS

The experimental results are compiled in Appendix A and Appendix B. These results were studied in an attempt to pinpoint the characteristics of the methods used and any significant differences between them. The following paragraphs contain detailed discussions of the experimental results and, where possible, rankings of the tests based on the several criteria mentioned earlier. It should be remembered that these results pertain to very specific configurations of the population means. Further, the results can only be considered approximately correct for the cases examined since they were subject to statistical variation. For example, the standard deviation of the experimentwise error rate of an approximately five percent test was nearly 0.01, and for a one-percent test it was about .0045. In spite of these problems, the results did indicate some obvious differences among the methods and permitted some fairly general conclusions.

A. UNCONDITIONAL COMPARISONS

The results of the unconditional experiments are presented by category in the following paragraphs.

1. Type I Errors Under H_o

The estimated per-comparison and experimentwise error rates when the null hypothesis was true are displayed in Table I. The experimental results indicated that these

error rates were independent of the sample size, and the values shown in the table were obtained by averaging the rates obtained for the six different sample sizes. The estimated per-comparison error rates for all but the Duncan procedure decreased with increasing values of k. Among those for the Duncan procedure the trend was less clear, but the per-comparison error rates clearly did not increase with increasing values of k. This trend is what one should expect, considering the general philosophy of multiple comparisons. Independent of the significance level of the tests and the number of means being tested, the ordering of test procedures from low to high based upon per-comparison error rates under H_o was: Scheffé < Tukey < S-N-K < Duncan.

The experimentwise error rate for the Duncan tests increased rapidly with increasing k, and for k = 5 it had reached 0.195 using a five percent test. For the other tests no clear relationship with k presented itself. As before an ordering, independent of the significance level and k, was possible. The ordering from low to high based upon experimentwise error rates was:

Scheffé < Tukey = S-N-K < Duncan

Note that although the Tukey and S-N-K procedures yielded different per-comparison error rates, their experimentwise error rates were identical. The Duncan procedure was designed to place primary emphasis on the control of the per-comparison error rate, and the experimental results

point out a relative lack of control of the experimentwise error rate. The Scheffé procedure appeared in difficulty in the opposite direction. One should have expected a five percent test to yield an experimentwise error rate of about five percent. The experimental results showed that the experimentwise error rate was significantly lower, indicating the procedure is overly conservative for contrasts of this type.

2. Type I Errors under the Alternate Hypothesis

The per-comparison and experimentwise error rates when the null hypothesis was false are given in Tables II and III. Table II shows the rates for the Scheffé and Tukey procedures, which did not depend upon d. The error rates for the two multiple range techniques did indicate a dependence on d, and these results are displayed in Table III. This difference arises because the credo of the multiple range tests requires that two means cannot be declared significant unless every subgroup of the k means which contains them is declared significant and because different critical values are used to test groups of means of different size. The dependence on d was much more pronounced for the S-N-K method than for that of Duncan.

The results showed that the per-comparison error rates for all but the Duncan procedure decreased with increasing k. The dependence in the case of the Duncan method was not clear, but it was not increasing. The ranking from low to high of the procedures for per-comparison error rates was: Scheffé < Tukey < S-N-K < Duncan

The estimated experimentwise error rates when H_o was false showed no obvious dependence on k for the Scheffé and S-N-K procedures, slightly increasing rates for the Tukey method, and a pronounced increasing relationship for the Duncan test. Within the range of k values considered, all tests except the Duncan test maintained their experimentwise error rates near or below the labeled significance level of the test; the Duncan procedure demonstrated little control over this type of error. The ranking from high to low based on the experimentwise error rate was:

Scheffé < Tukey < S-N-K < Duncan

3. Power

Appendix B contains a series of plots of estimated power as a function of the true difference d and the sample size n. All of the procedures showed increasing power as the sample size increased. Power of the Scheffé and Tukey procedures decreased as the number of population means being tested was increased. The S-N-K procedure showed a similar, but less pronounced, decrease in power; however the Duncan procedure showed little, if any, decrease in power as k increased. For all of the cases studied there was a clear ordering of the tests based on power in detecting differences between pairs of means. The ordering from least powerful to most powerful was:

Scheffé < Tukey < S-N-K < Duncan

B. CONDITIONAL COMPARISONS

In many respects the results of the conditional comparisons were similar to those of the unconditional experiment. The changes noted occurred primarily when the null hypothesis was true and for small values of d. As d became large the results approached those of the unconditioned experiments. This was not surprising since for d large enough one would expect the two experiments to be providing identical sets of means for test by the multiple comparison procedures. The following paragraphs point out the characteristic changes induced by the conditioning process. Even though not specifically pointed out in every case, it should be kept in mind that these differences diminished with increasing values of d.

1. Type I Errors Under H_o

The orderings of the multiple comparison techniques based on experimentwise and per-comparison error rates obtained for the unconditional experiment were not changed. The magnitudes of the error rates increased greatly as a result of the conditioning process as was expected. To answer questions about the advisiability or necessity of first testing for equality of means, the law of total probability was used to calculate the difference in overall error rates resulting from conditioning. The correct probability statement was:

$$\Pr\{\text{Type I error by a MC procedure} \mid H_o\}$$

$$= \Pr\{\text{Type I error by a MC procedure} \mid \text{Type I error by ANOV procedure}, H_o\} \times \Pr\{\text{Type I error by ANOV procedure} \mid H_o\} + \Pr\{\text{Type I error by MC procedure} \mid \text{ANOV procedure correct}, H_o\} \times \Pr\{\text{ANOV procedure correct} \mid H_o\}$$

$$= \Pr\{\text{Type I by MC} \mid \text{Type I by ANOV}\}\ \alpha + \Pr\{\text{Type I by} \mid \text{MC ANOV currect}\}\ (1-\alpha)$$

where $\alpha = \Pr\{\text{Type I error by ANOV} \mid H_o\}$.

The quantity on the left hand side was estimated for each test procedure from the results of the unconditional experiments. The results of the conditional experiments gave the probabilities of a Type I error by the multiple comparison methods given that a Type I error was made in the analysis of variance procedure under H_o. The nominal significance level of the analysis of variance used in the experiment was $\alpha = .05$. Data were not collected to evaluate the second term on the right-hand side because of the large number of tests which would have been required. The two known quantities were calculated from the experimental results and compared. These calculations showed that for both kinds of Type I errors by the Scheffé, Tukey, and S-N-K methods the known quantities were nearly equal in all cases. Accordingly, the probability of a Type I error by these methods when the analysis of variance correctly accepts H_o must be very small, possibly zero. This result indicated that for these methods, the prior performance of an analysis of variance offered little or no additional protection against a Type I error of either kind. It did not mean,

however, that the analysis of variance and these multiple comparison methods were equivalent since the estimated experimentwise error rates of the multiple comparison procedures were all less than one. In other words there were trials on which the analysis of variance incorrectly rejected the null hypothesis, but the multiple comparison procedures did not reject.

In the case of Duncan's procedure, the experimentwise error rate was one or nearly one for all values of k used in the experiment, indicating that whenever the analysis of variance incorrectly rejected H_o, the Duncan test also rejected. Further, the unconditional probabilities of Type I errors were significantly larger than the calculated values of the overall Type I errors resulting from conditioning. In contrast to the results for the other three methods, this indicated that the prior performance of an analysis of variance did offer significantly increased protection against Type I errors under the null hypothesis and maintained the experimentwise error rate at or near the nominal significance level of the test.

2. Type I Errors Under the Alternate Hypothesis

When the null hypothesis was false both kinds of conditional Type I error rates were higher than the unconditional Type I error rates for all cases studied. In contrast with the unconditional rates, the rates for the conditional experiment were not independent of the sample size. The rates decreased as the sample size increased - the

decrease being more pronounced for smaller values of d. Contrary to the previous results, both kinds of Type I error rate decreased with increasing values of d, and they approached the unconditional rates for large values of d. The conditional per-comparison error rates decreased with increasing values of k for small values of d and showed the same mixed tendencies as the unconditional rates for large d. The conditional experimentwise error rate for the Scheffé procedure decreased with k; all other procedures displayed increasing rates as k increased, slowly increasing for the Tukey and S-N-K methods, and rapidly so for Duncan's procedure. The results showed in general that when d was small, there was a fairly high probability that the differences declared significant by the multiple comparison procedures would be the wrong ones.

3. Power

The conditional power figures, shown in Appendix B, were greater than those for the unconditional experiment as was expected. The greatest increase was for small values of d and the differences between conditional and unconditional power decreased with increasing d. The conditional power curves were not as smooth as the curves of unconditional power. This lack of smoothness was probably caused by the method of selection of data sets to be used for the multiple comparisons tests. The irregularities resulted from the statistical variation between the data tested for two different values of d. In the unconditional

experiments this variation was eliminated by using the same sets of data for all values of d. All other characteristics of the power function remained unchanged under conditioning, and the ordering of test methods based on power was unchanged.

C. CONTRASTS OF THREE MEANS

The final experiment, consisting of tests of contrasts of three means, showed that for this type of contrast the Tukey procedure yielded smaller Type I error rates than the Scheffé procedure. The Type I error rates appeared to be independent of the sample size and were averaged to obtain the estimated error rates. The estimated per-comparison error rates of the five percent significance level tests were .0153 for the Scheffé method and .0077 for the Tukey method. The estimated experimentwise error rates of the five percent tests were .0393 for Scheffé and .0207 for Tukey.

The curves of estimated power plotted in Figures 10 and 20 showed that as predicted the Scheffé test was more powerful than that of Tukey for this type of contrast.

VI. CONCLUSIONS

Although the experiments were conducted using only one of the many possible ways in which a group of several population means could differ, the results obtained should be applicable for other cases as well. Scheffé's procedure seemed far too conservative when used to test comparisons of only two means. This result was not unexpected, since the method is completely general whereas the others were designed more specifically for contrasts of this type. On the other hand, in cases where blends or mixtures could be important, the Scheffé procedure would be a better choice than that of Tukey.

The appropriate choice among the remaining three methods for use when only contrasts of two means are important seemed to depend upon the relative importance the experimenter attaches to the two different kinds of Type I errors. It was concluded that the experimenter whose primary concern is control of the per-comparison error rate and is not worried about the experimentwise error rate should use Duncan's method, but only after first testing for equality of the means with an analysis of variance procedure.

For the true multiple comparisonist who desires to control the experimentwise error rate, the choice lies between the S-N-K method and that of Tukey. The two procedures have identical experimentwise error rates under the null hypothesis, but power, per-comparison error rates, and

experimentwise error rates under the alternate hypothesis all differ. The S-N-K procedure maintained the experimentwise error rate under the alternate hypothesis near the significance level of the test, whereas the Tukey method was more conservative. The over conservatism of the Tukey method resulted in lower power for all values of d and n, and consequently the S-N-K method appeared to be the better choice.

Figure 39 shows the estimated power of all four methods side by side for a typical case to aid visualization of the magnitude of power differential involved.

APPENDIX A

TABLE I

ESTIMATED PER-COMPARISON (PC) AND EXPERIMENTWISE (E) ERROR RATES UNDER THE NULL HYPOTHESIS

Test	Experiment	%	Type	k=3	k=4	k=5
Scheffe	uncond.	1	PC	.0025	.0011	.0006
Tukey	uncond.	1	PC	.0030	.0019	.0017
S-N-K	uncond.	1	PC	.0040	.0025	.0021
Duncan	uncond.	1	PC	.0081	.0057	.0066
Scheffe	uncond.	5	PC	.0148	.0049	.0028
Tukey	uncond.	5	PC	.0190	.0089	.0073
S-N-K	uncond.	5	PC	.0236	.0114	.0092
Duncan	uncond.	5	PC	.0428	.0368	.0356
Scheffe	uncond.	1	E	.0070	.0053	.0057
Tukey	uncond.	1	E	.0097	.0100	.0140
S-N-K	uncond.	1	E	.0097	.0100	.0140
Duncan	uncond.	1	E	.0200	.0250	.0410
Scheffe	uncond.	5	E	.0373	.0243	.0213
Tukey	uncond.	5	E	.0480	.0417	.0507
S-N-K	uncond.	5	E	.0480	.0417	.0507
Duncan	uncond.	5	E	.0977	.1443	.1950
Scheffe	cond.	5	PC	.2965	.1056	.0456
Tukey	cond.	5	PC	.3549	.1813	.1114
S-N-K	cond.	5	PC	.4474	.2420	.1490
Duncan	cond.	5	PC	.5155	.3604	.2860
Scheffe	cond.	5	E	.6497	.5363	.3823
Tukey	cond.	5	E	.8910	.8357	.7913
S-N-K	cond.	5	E	.8910	.8357	.7913
Duncan	cond.	5	E	1.0000	.9990	.9997

TABLE II

ESTIMATED PER-COMPARISON (PC) AND EXPERIMENTWISE (E) ERROR RATES FOR THE TUKEY AND SCHEFFE PROCFDURES UNDER THE ALTERNATE HYPOTHESIS IN THE UNCONDITIONAL EXPERIMENTS

Test	Level	Type	k=3	k=4	k=5
Scheffé	1%	PC	.0023	.0011	.0008
Tukey	1	PC	.0037	.0019	.0019
Scheffé	5	PC	.0170	.0042	.0030
Tukey	5	PC	.0206	.0086	.0077
Scheffé	1	E	.0023	.0026	.0043
Tukey	1	E	.0037	.0050	.0097
Scheffé	5	F	.0170	.0113	.0143
Tukey	5	E	.0206	.0230	.0343

TABLE III

ESTIMATED PER-COMPARISON (PC) AND EXPERIMENTWISE (E) ERROR RATES UNDER THE ALTERNATE HYPOTHESIS FOR MULTIPLE RANGE PROCEDURES IN UNCONDITIONAL EXPERIMENTS

Test	Type	k	%	d=0.2	d=0.4	d=0.6	d=0.8	d=1.0	d=1.2	d=1.4	d=1.6	d=1.8	d=2.0
S-N-K	PC	3	5	.0283	.0350	.0427	.0480	.0500	.0510	.0527	.0533	.0533	.0537
	PC	4	5	.0130	.0154	.0170	.0192	.0206	.0212	.0213	.0213	.0216	.0217
	PC	5	5	.0106	.0119	.0128	.0136	.0141	.0145	.0148	.0150	.0151	.0151
	PC	3	1	.0057	.0073	.0087	.0097	.0103	.0107	.0117	.0117	.0120	.0120
	PC	4	1	.0024	.0029	.0029	.0030	.0031	.0031	.0032	.0033	.0033	.0033
	PC	5	1	.0021	.0023	.0026	.0027	.0027	.0030	.0030	.0031	.0031	.0031
Duncan	PC	3	5	.0480	.0500	.0513	.0527	.0530	.0530	.0533	.0533	.0537	.0537
	PC	4	5	.0368	.0386	.0391	.0398	.0398	.0400	.0400	.0402	.0404	.0404
	PC	5	5	.0381	.0386	.0389	.0399	.0402	.0403	.0405	.0406	.0406	.0406
	PC	3	1	.0090	.0103	.0110	.0110	.0110	.0117	.0117	.0117	.0120	.0120
	PC	4	1	.0054	.0056	.0059	.0061	.0061	.0061	.0061	.0061	.0062	.0062
	PC	5	1	.0066	.0068	.0069	.0070	.0071	.0074	.0074	.0074	.0074	.0074
S-N-K	E	3	5	.0283	.0350	.0427	.0480	.0500	.0510	.0527	.0533	.0533	.0537
	E	4	5	.0283	.0333	.0383	.0403	.0427	.0443	.0443	.0443	.0450	.0450
	E	5	5	.0377	.0427	.0470	.0490	.0503	.0510	.0517	.0517	.0517	.0517
	E	3	1	.0057	.0073	.0087	.0097	.0103	.0107	.0117	.0117	.0120	.0120
	E	4	1	.0053	.0067	.0067	.0073	.0073	.0073	.0077	.0077	.0077	.0077
	E	5	1	.0097	.0103	.0110	.0120	.0120	.0123	.0123	.0127	.0127	.0127
Duncan	E	3	5	.0480	.0500	.0513	.0527	.0530	.0530	.0533	.0533	.0537	.0537
	E	4	5	.0860	.0883	.0893	.0907	.0907	.0913	.0913	.0917	.0917	.0917
	E	5	5	.1440	.1447	.1470	.1480	.1487	.1490	.1500	.1503	.1503	.1503
	E	3	1	.0090	.0103	.0110	.0110	.0110	.0117	.0117	.0117	.0120	.0120
	E	4	1	.0140	.0143	.0153	.0157	.0157	.0157	.0157	.0157	.0160	.0160
	E	5	1	.0273	.0277	.0283	.0290	.0293	.0300	.0300	.0300	.0300	.0300

TABLE IV

ESTIMATED PER-COMPARISON/EXPERIMENTWISE ERROR RATES UNDER THE ALTERNATIVE HYPOTHESIS IN THE CONDITIONAL EXPERIMENTS

Test	k	n	d=0.2	d=0.4	d=0.6	d=0.8	d=1.0	d=1.2	d=1.4	d=1.6	d=1.8	d=2.0
Scheffé	3	6	.3360	.1767	.0999	.0720	.0540	.0399	.0399	.0240	.0201	.0180
	3	8	.2721	.1320	.0759	.0501	.0381	.0279	.0261	.0240	.0240	.0240
	3	10	.2421	.1119	.0639	.0261	.0120	.0099	.0081	.0081	.0081	.0081
	3	20	.1680	.0621	.0279	.0180	.0159	.0141	.0141	.0141	.0141	.0141
	3	30	.1239	.0339	.0240	.0180	.0120	.0120	.0120	.0120	.0120	.0120
	3	40	.0840	.0381	.0219	.0201	.0201	.0201	.0201	.0201	.0201	.0201
Tukey	3	6	.3939	.2079	.1200	.0840	.0621	.0441	.0381	.0279	.0219	.0211
	3	8	.3441	.1701	.1020	.0639	.0480	.0360	.0321	.0300	.0300	.0300
	3	10	.3039	.1479	.0879	.0360	.0180	.0141	.0120	.0099	.0099	.0081
	3	20	.2139	.0900	.0441	.0261	.0231	.0180	.0180	.0180	.0180	.0180
	3	30	.1560	.0459	.0261	.0211	.0141	.0141	.0141	.0141	.0141	.0141
	3	40	.1161	.0459	.0261	.0240	.0240	.0240	.0240	.0240	.0240	.0240
S-N-K	3	6	.3820	.2520	.1780	.1280	.1100	.0820	.0780	.0680	.0620	.0560
	3	8	.3280	.2440	.1880	.1240	.0940	.0740	.0620	.0540	.0540	.0540
	3	10	.3140	.2020	.1460	.0900	.0540	.0440	.0380	.0340	.0340	.0320
	3	20	.2660	.1280	.0760	.0660	.0460	.0440	.0420	.0420	.0420	.0420
	3	30	.1960	.0860	.0620	.0440	.0380	.0380	.0380	.0380	.0380	.0380
	3	40	.1620	.1020	.0660	.0680	.0680	.0680	.0680	.0680	.0680	.0680
Duncan	3	6	.4420	.2960	.2180	.1600	.1240	.0920	.0820	.0680	.0620	.0560
	3	8	.3920	.2840	.2040	.1320	.0980	.0740	.0620	.0540	.0540	.0540
	3	10	.3660	.2360	.1680	.0980	.0560	.0440	.0380	.0340	.0340	.0320
	3	20	.3180	.1520	.0880	.0700	.0460	.0440	.0420	.0420	.0420	.0420
	3	30	.2520	.1080	.0660	.0440	.0380	.0380	.0380	.0380	.0380	.0380
	3	40	.2080	.1100	.0700	.0680	.0680	.0680	.0680	.0680	.0680	.0680

TABLE V

ESTIMATED PER-COMPARISON FRROR RATES UNDER THE ALTFRNATIVE HYPOTHESIS IN THF CONDITIONAL EXPFRIMENTS

Test	k	n	d=0.2	d=0.4	d=0.6	d=0.8	d=1.0	d=1.2	d=1.4	d=1.6	d=1.8	d=2.0
Scheffé	4	6	.0980	.0553	.0340	.0227	.0180	.0120	.0073	.0067	.0067	.0067
	4	8	.0827	.0413	.0233	.0173	.0107	.0073	.0060	.0060	.0060	.0060
	4	10	.0627	.0327	.0187	.0113	.0080	.0060	.0047	.0027	.0027	.0027
	4	20	.0593	.0233	.0093	.0060	.0053	.0040	.0040	.0040	.0040	.0040
	4	30	.0393	.0167	.0127	.0093	.0087	.0087	.0087	.0087	.0087	.0087
	4	40	.0373	.0147	.0080	.0047	.0047	.0047	.0047	.0047	.0047	.0047
Tukey	4	6	.1640	.1007	.0660	.0433	.0313	.0233	.0167	.0140	.0133	.0133
	4	8	.1480	.0793	.0500	.0333	.0227	.0153	.0127	.0113	.0113	.0113
	4	10	.1380	.0747	.0420	.0273	.0180	.0147	.0120	.0100	.0093	.0093
	4	20	.1160	.0467	.0227	.0147	.0127	.0107	.0107	.0107	.0107	.0107
	4	30	.0767	.0320	.0187	.0133	.0120	.0120	.0120	.0120	.0120	.0120
	4	40	.0707	.0287	.0167	.0100	.0093	.0093	.0093	.0093	.0093	.0093
S-N-K	4	6	.1880	.1427	.1040	.0833	.0680	.0520	.0413	.0347	.0313	.0307
	4	8	.1807	.1240	.0900	.0667	.0473	.0340	.0267	.0240	.0240	.0233
	4	10	.1727	.1113	.0680	.0500	.0353	.0273	.0227	.0200	.0187	.0187
	4	20	.1253	.0647	.0480	.0360	.0300	.0260	.0260	.0260	.0260	.0260
	4	30	.1167	.0613	.0413	.0313	.0287	.0287	.0287	.0287	.0287	.0287
	4	40	.0947	.0420	.0247	.0200	.0187	.0187	.0187	.0187	.0187	.0187
Duncan	4	6	.3380	.2560	.1873	.1393	.1113	.0833	.0687	.0580	.0507	.0473
	4	8	.2913	.2107	.1653	.1173	.0813	.0560	.0420	.0360	.0353	.0340
	4	10	.2706	.1860	.1220	.0947	.0673	.0573	.0487	.0447	.0433	.0433
	4	20	.2293	.1260	.0833	.0613	.0500	.0440	.0433	.0433	.0433	.0433
	4	30	.2033	.1020	.0580	.0440	.0413	.0413	.0413	.0413	.0413	.0413
	4	40	.1733	.0767	.0480	.0380	.0353	.0353	.0353	.0353	.0353	.0353

TABLE VI

ESTIMATED EXPERIMENTWISE ERROR RATES UNDER THE ALTERNATIVE HYPOTHESIS IN THE CONDITIONAL EXPERIMENTS

Test	k	n	d=0.2	d=0.4	d=0.6	d=0.8	d=1.0	d=1.2	d=1.4	d=1.6	d=1.8	d=2.0
Scheffé	4	6	.2580	.1440	.0880	.0600	.0460	.0280	.0180	.0160	.0160	.0160
	4	8	.2240	.1140	.0680	.0500	.0300	.0200	.0160	.0160	.0160	.0160
	4	10	.1700	.0900	.0540	.0320	.0220	.0160	.0120	.0080	.0080	.0080
	4	20	.1660	.0620	.0260	.0180	.0160	.0120	.0120	.0120	.0120	.0120
	4	30	.1060	.0460	.0340	.0260	.0240	.0240	.0240	.0240	.0240	.0240
	4	40	.1060	.0420	.0220	.0120	.0120	.0120	.0120	.0120	.0120	.0120
Tukey	4	6	.4120	.2500	.1620	.1120	.0820	.0580	.0420	.0340	.0320	.0320
	4	8	.3760	.2080	.1400	.0940	.0620	.0420	.0340	.0300	.0300	.0300
	4	10	.3660	.2020	.1140	.0740	.0500	.0400	.0320	.0280	.0260	.0260
	4	20	.3140	.1280	.0640	.0420	.0360	.0300	.0300	.0300	.0300	.0300
	4	30	.1980	.0840	.0480	.0360	.0320	.0320	.0320	.0320	.0320	.0320
	4	40	.1860	.0720	.0420	.0280	.0260	.0260	.0260	.0260	.0260	.0260
S-N-K	4	6	.3920	.2980	.2140	.1660	.1340	.1020	.0820	.0700	.0620	.0600
	4	8	.3820	.2620	.1980	.1500	.1060	.0760	.0600	.0540	.0540	.0520
	4	10	.3900	.2440	.1480	.1100	.0780	.0600	.0480	.0440	.0420	.0420
	4	20	.2800	.1400	.1000	.0760	.0680	.0580	.0580	.0580	.0580	.0580
	4	30	.2660	.1360	.0840	.0620	.0540	.0540	.0540	.0540	.0540	.0540
	4	40	.2060	.0980	.0520	.0420	.0400	.0400	.0400	.0400	.0400	.0400
Duncan	4	6	.6780	.5360	.3840	.2880	.2360	.1800	.1540	.1300	.1100	.1020
	4	8	.5920	.4680	.3760	.2700	.1920	.1320	.1000	.0860	.0840	.0800
	4	10	.5720	.4080	.2780	.2220	.1640	.1400	.1180	.1100	.1080	.1080
	4	20	.4840	.2780	.1820	.1400	.1180	.1040	.1020	.1020	.1020	.1020
	4	30	.4540	.2300	.1280	.0980	.0900	.0900	.0900	.0900	.0900	.0900
	4	40	.3800	.1800	.1180	.0940	.0900	.0900	.0900	.0900	.0900	.0900

TABLE VII

ESTIMATED PER-COMPARISON ERROR RATES UNDER THE ALTERNATIVE HYPOTHESIS IN THE CONDITIONAL EXPERIMENTS

Test	k	n	d=0.2	d=0.4	d=0.6	d=0.8	d=1.0	d=1.2	d=1.4	d=1.6	d=1.8	d=2.0
Scheffé	5	6	.0480	.0346	.0240	.0174	.0146	.0094	.0056	.0046	.0044	.0040
	5	8	.0310	.0174	.0110	.0066	.0040	.0030	.0020	.0014	.0014	.0014
	5	10	.0386	.0190	.0104	.0070	.0036	.0026	.0026	.0024	.0024	.0024
	5	20	.0206	.0080	.0026	.0016	.0010	.0010	.0010	.0010	.0010	.0010
	5	30	.0186	.0066	.0044	.0040	.0033	.0033	.0033	.0033	.0033	.0033
	5	40	.0130	.0046	.0026	.0024	.0024	.0024	.0024	.0024	.0024	.0024
Tukey	5	6	.1116	.0776	.0570	.0414	.0314	.0224	.0160	.0136	.0124	.0114
	5	8	.0874	.0520	.0344	.0206	.0134	.0084	.0056	.0044	.0044	.0040
	5	10	.0880	.0480	.0280	.0200	.0134	.0100	.0100	.0086	.0086	.0086
	5	20	.0644	.0246	.0126	.0086	.0074	.0074	.0074	.0074	.0074	.0074
	5	30	.0516	.0204	.0116	.0096	.0090	.0083	.0083	.0083	.0083	.0083
	5	40	.0414	.0160	.0074	.0064	.0064	.0064	.0064	.0064	.0064	.0064
S-N-K	5	6	.1400	.1110	.0837	.0613	.0463	.0370	.0280	.0267	.0240	.0220
	5	8	.1100	.0733	.0537	.0377	.0280	.0187	.0140	.0113	.0110	.0107
	5	10	.1073	.0653	.0393	.0313	.0210	.0177	.0167	.0153	.0153	.0153
	5	20	.0923	.0493	.0307	.0213	.0170	.0167	.0167	.0167	.0167	.0167
	5	30	.0690	.0323	.0143	.0087	.0077	.0077	.0077	.0077	.0077	.0077
	5	40	.0657	.0220	.0133	.0080	.0080	.0080	.0080	.0080	.0080	.0080
Duncan	5	6	.2693	.2230	.1723	.1357	.1103	.0883	.0683	.0593	.0503	.0470
	5	8	.2340	.1747	.1377	.1033	.0743	.0547	.0440	.0370	.0340	.0320
	5	10	.2383	.1690	.1100	.0840	.0623	.0520	.0483	.0460	.0453	.0453
	5	20	.2033	.1260	.0810	.0593	.0487	.0457	.0457	.0457	.0457	.0457
	5	30	.1607	.0870	.0470	.0340	.0317	.0317	.0317	.0317	.0317	.0317
	5	40	.1513	01687	.0420	.0310	.0307	.0303	.0303	.0303	.0303	.0303

TABLE VIII

ESTIMATED EXPERIMENTWISE ERROR RATES UNDER THE ALTERNATIVE HYPOTHESIS IN THE CONDITIONAL EXPERIMENTS

Test	k	n	d=0.2	d=0.4	d=0.6	d=0.8	d=1.0	d=1.2	d=1.4	d=1.6	d=1.8	d=2.0
Scheffe	5	6	.2320	.1620	.1040	.0740	.0620	.0460	.0280	.0220	.0200	.0180
	5	8	.1680	.0960	.0640	.0380	.0220	.0180	.0120	.0080	.0080	.0080
	5	10	.2000	.1000	.0560	.0380	.0200	.0140	.0140	.0120	.0120	.0120
	5	20	.1140	.0440	.0140	.0080	.0040	.0040	.0040	.0040	.0040	.0040
	5	30	.0980	.0360	.0220	.0200	.0180	.0180	.0180	.0180	.0180	.0180
	5	40	.0660	.0240	.0140	.0120	.0120	.0120	.0120	.0120	.0120	.0120
Tukey	5	6	.5000	.3480	.2480	.1780	.1300	.0980	.0740	.0600	.0540	.0500
	5	8	.4120	.2480	.1700	.1060	.0720	.0440	.0300	.0260	.0260	.0240
	5	10	.4260	.2280	.1360	.0940	.0600	.0460	.0460	.0420	.0420	.0420
	5	20	.3100	.1300	.0660	.0460	.0380	.0380	.0380	.0380	.0380	.0380
	5	30	.2480	.0980	.0600	.0520	.0500	.0460	.0460	.0460	.0460	.0460
	5	40	.1960	.0740	.0360	.0300	.0300	.0300	.0300	.0300	.0300	.0300
S-N-K	5	6	.4980	.3880	.2840	.2100	.1540	.1240	.0980	.0840	.0780	.0720
	5	8	.4220	.2840	.2100	.1540	.1180	.0780	.0580	.0500	.0480	.0460
	5	10	.4320	.2560	.1640	.1260	.0840	.0680	.0660	.0620	.0620	.0620
	5	20	.3480	.1840	.1140	.0800	.0640	.0620	.0620	.0620	.0620	.0620
	5	30	.2720	.1260	.0600	.0340	.0320	.0320	.0320	.0320	.0320	.0320
	5	40	.2800	.0980	.0540	.0380	.0380	.0380	.0380	.0380	.0380	.0380
Duncan	5	6	.7900	.6880	.5480	.4440	.3660	.3080	.2380	.2080	.1780	.1700
	5	8	.7220	.5860	.4880	.3820	.2780	.2120	.1700	.1460	.1360	.1260
	5	10	.7580	.5800	.3960	.3100	.2420	.2060	.1900	.1800	.1760	.1760
	5	20	.6420	.4400	.3040	.2280	.1800	.1660	.1660	.1660	.1660	.1660
	5	30	.5420	.3100	.1840	.1280	.1220	.1220	.1220	.1220	.1220	.1220
	5	40	.5200	.2720	.1660	.1280	.1280	.1260	.1240	.1240	.1240	.1240

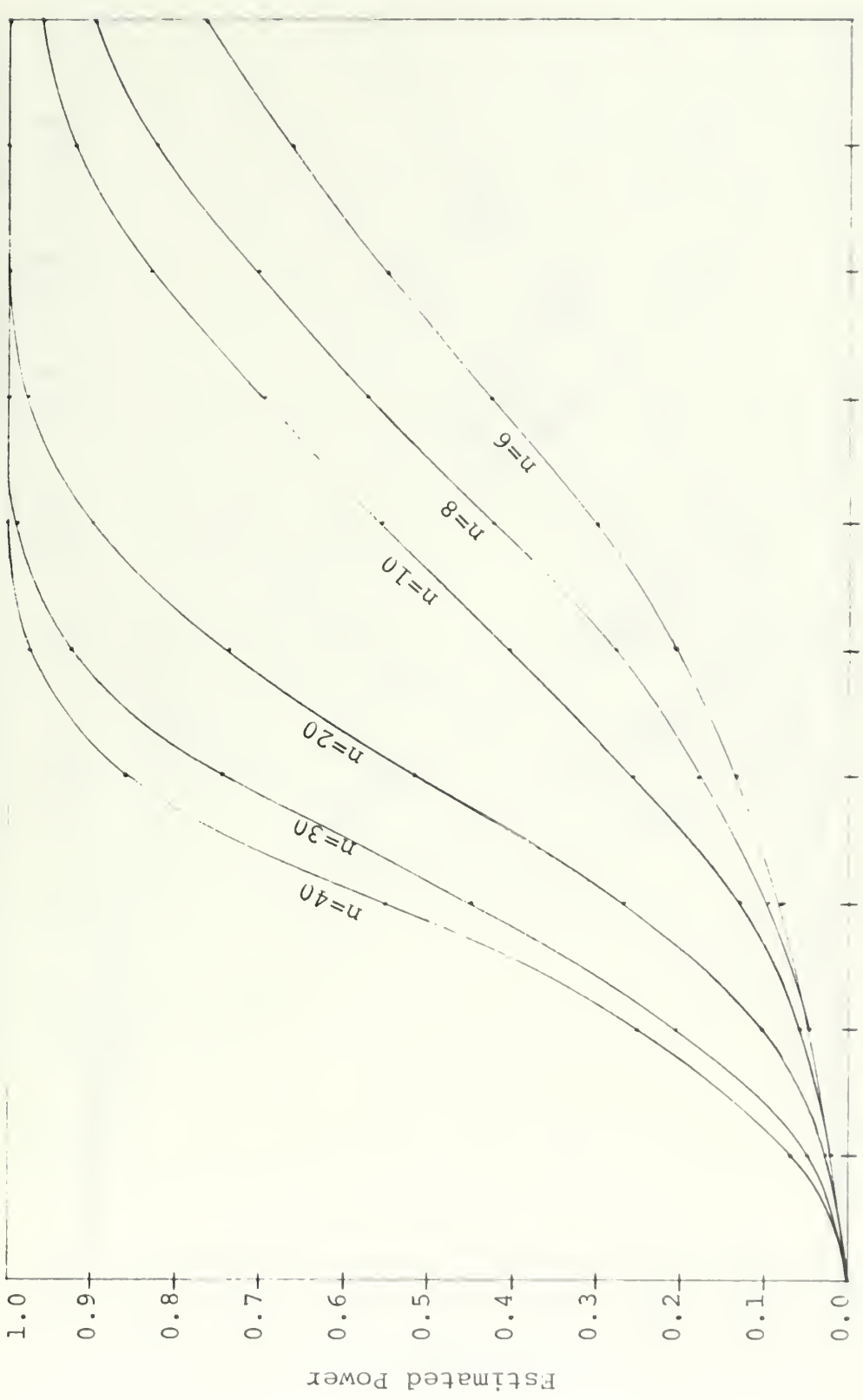

Figure 1. Estimated (unconditioned) power of 5% Scheffé test of three means

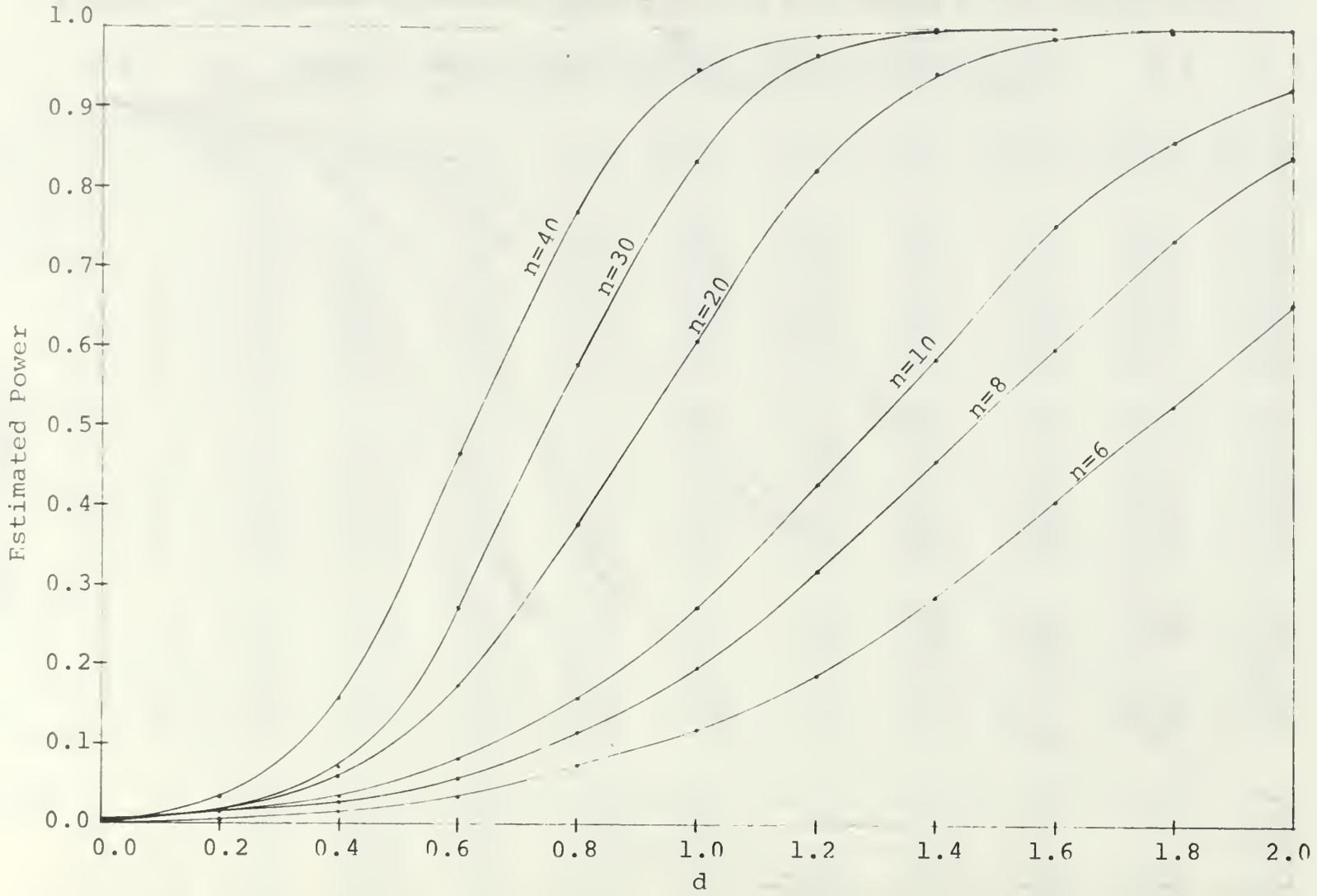

Figure 2. Estimated (unconditioned) power of 5% Scheffé test of four means

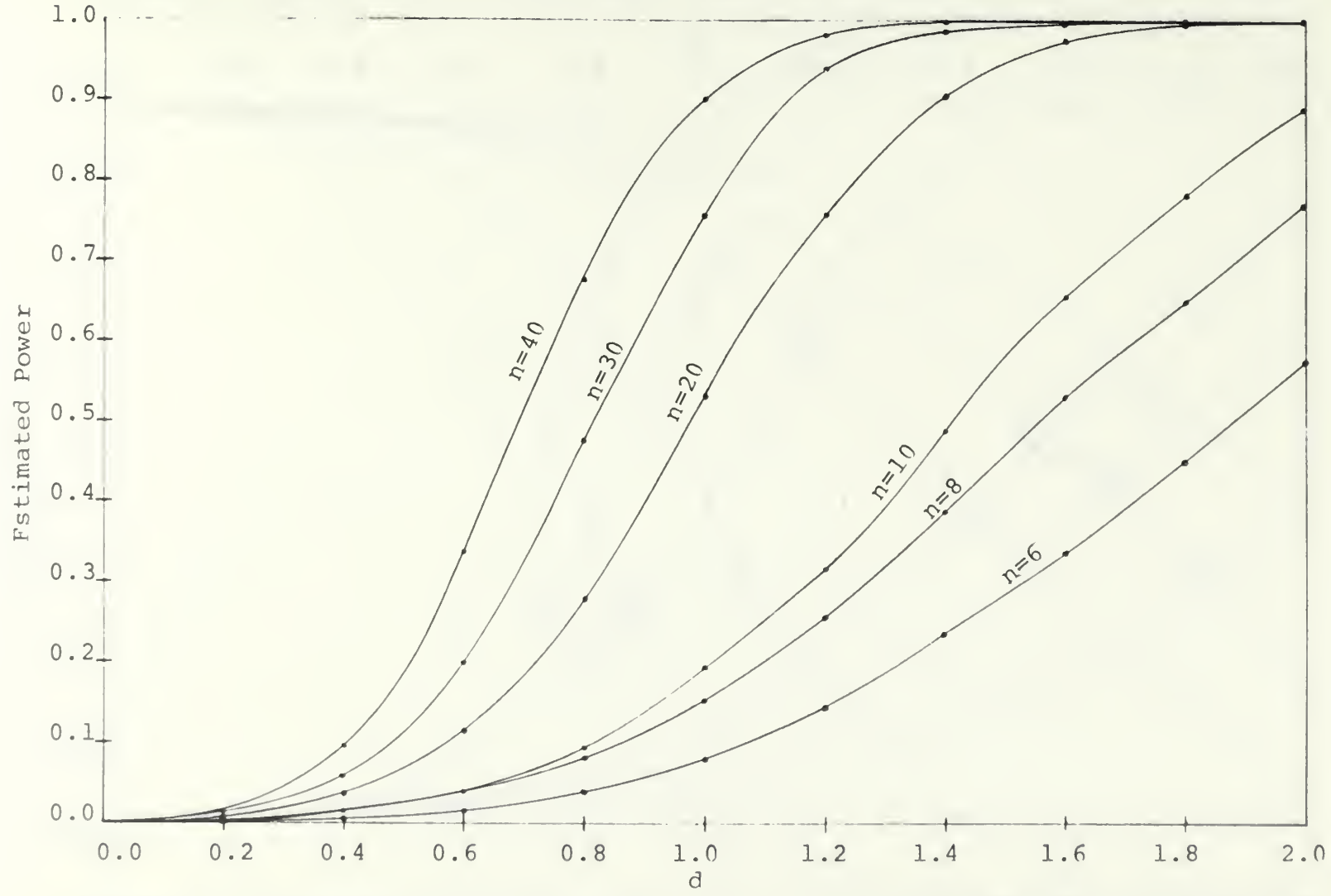

Figure 3. Estimated (unconditioned) power of 5% Scheffé test of five means

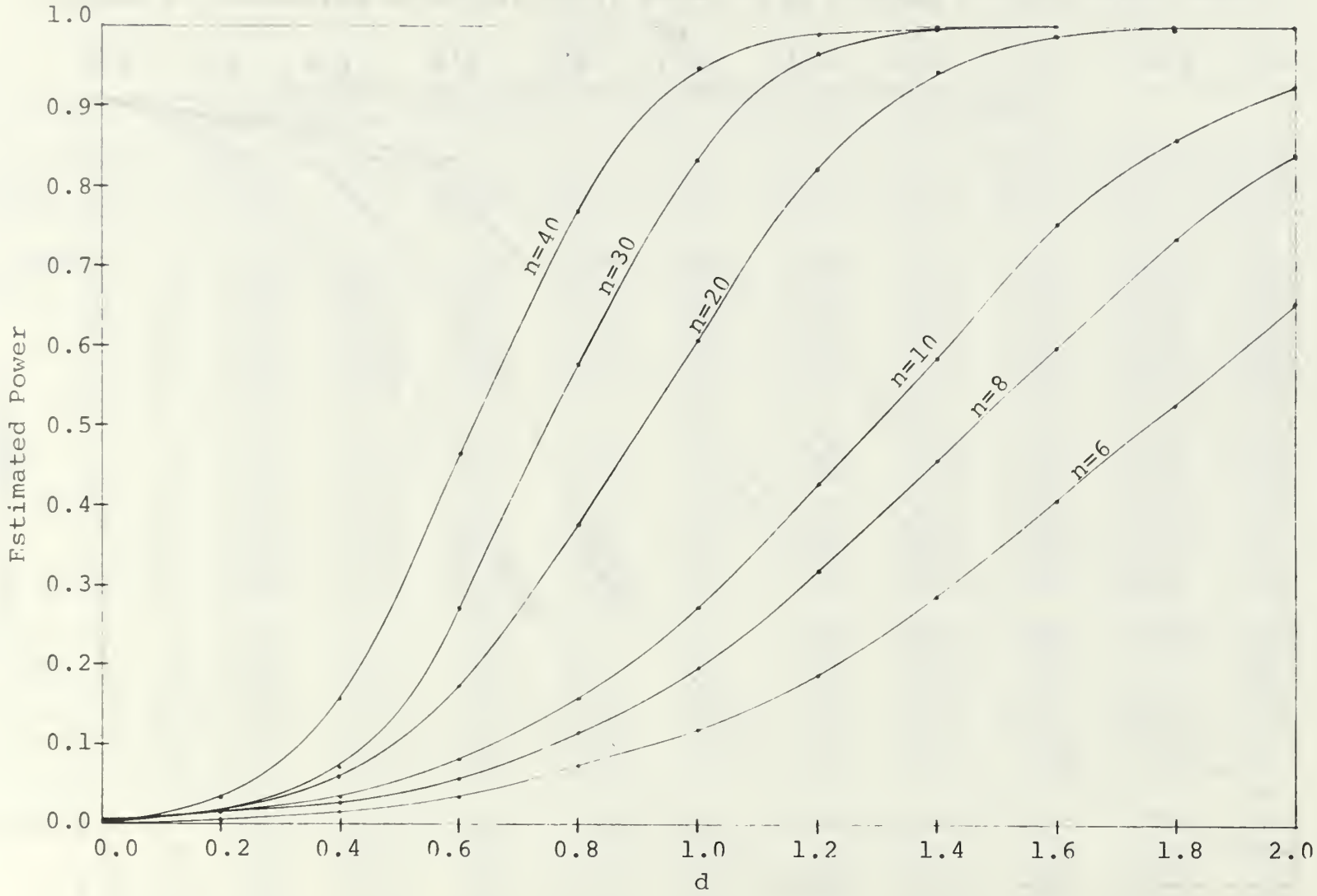

Figure 2. Estimated (unconditioned) power of 5% Scheffé test of four means

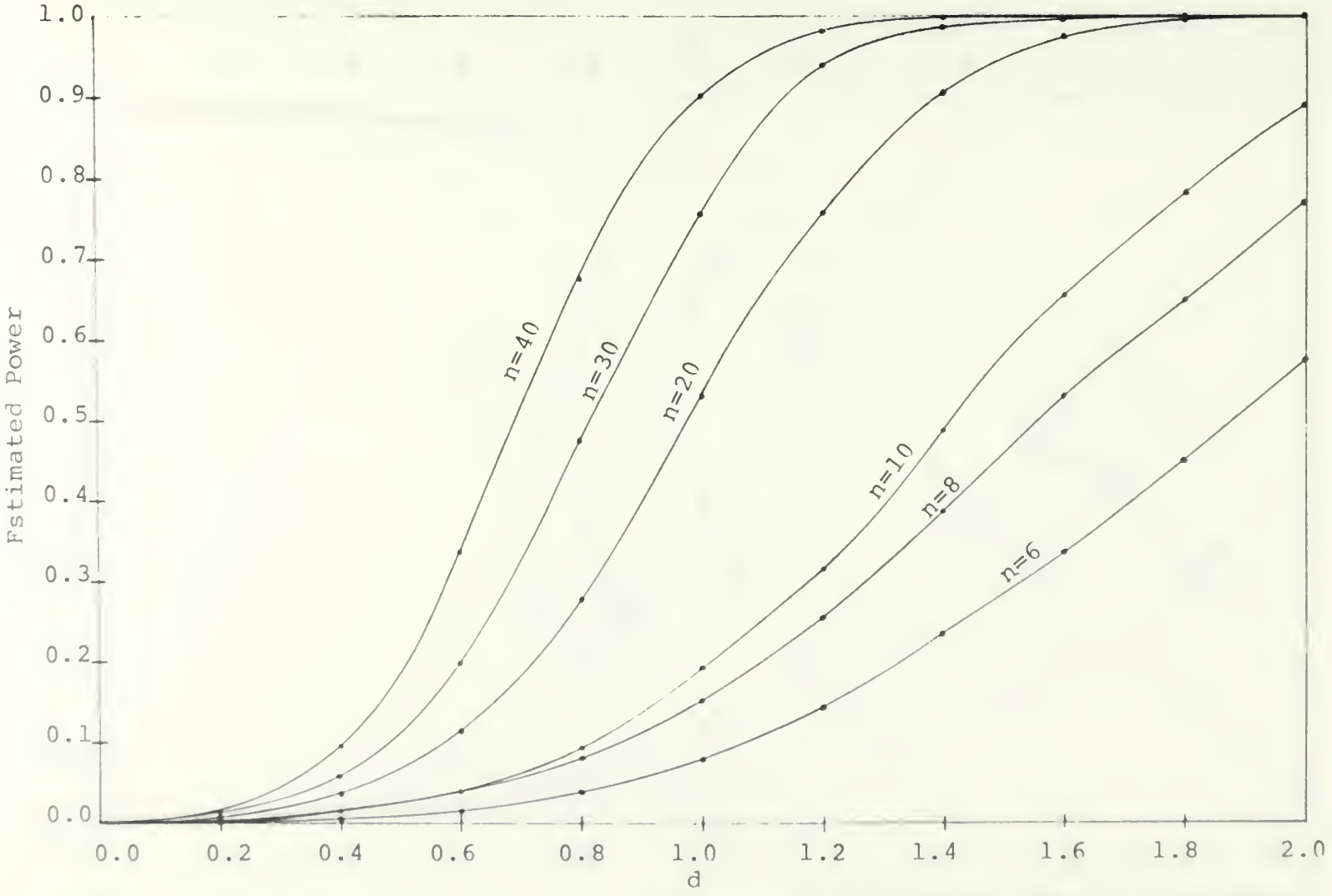

Figure 3. Estimated (unconditioned) power of 5% Scheffé test of five means

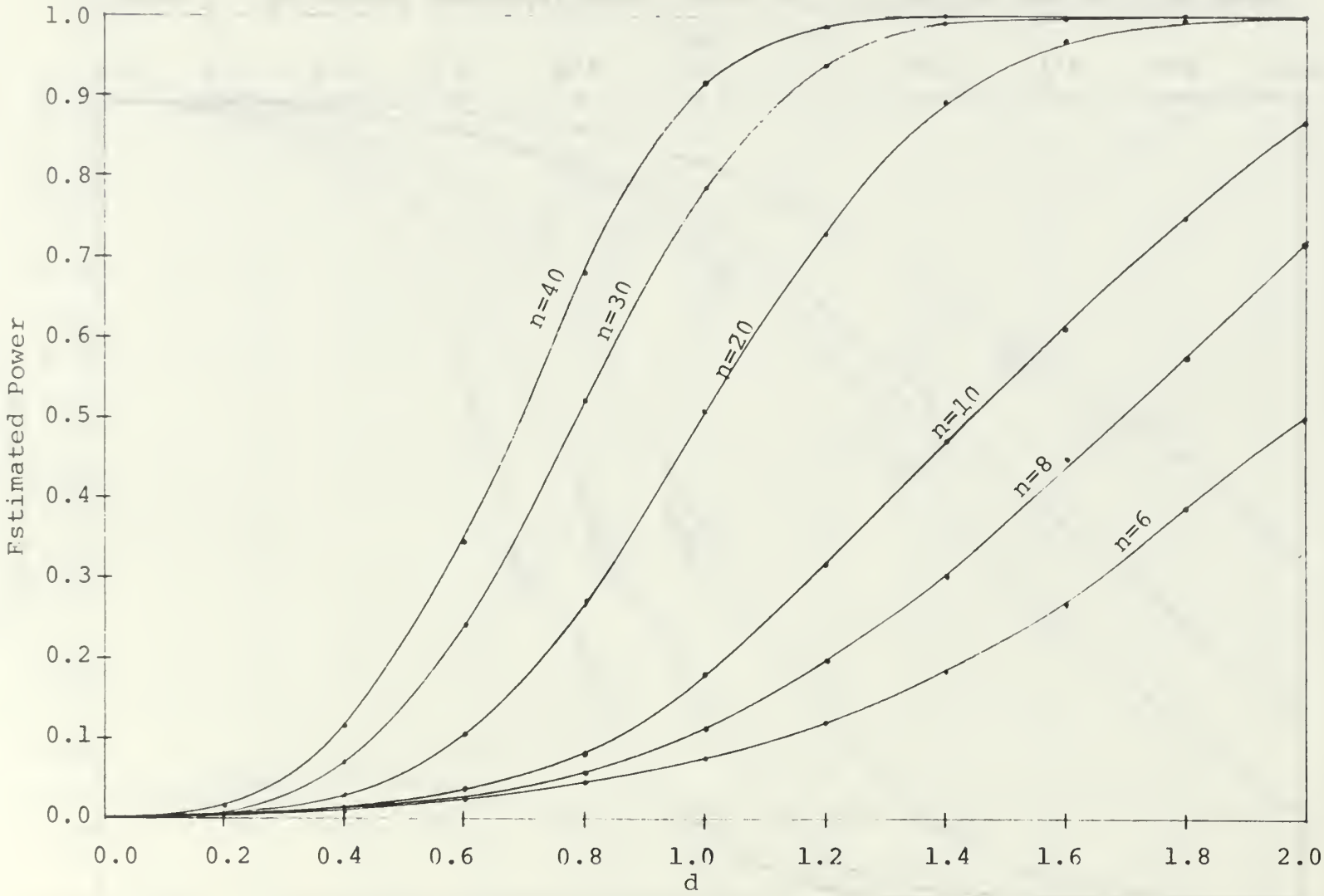

Figure 4. Estimated (unconditioned) power of 1% Scheffé test of three means

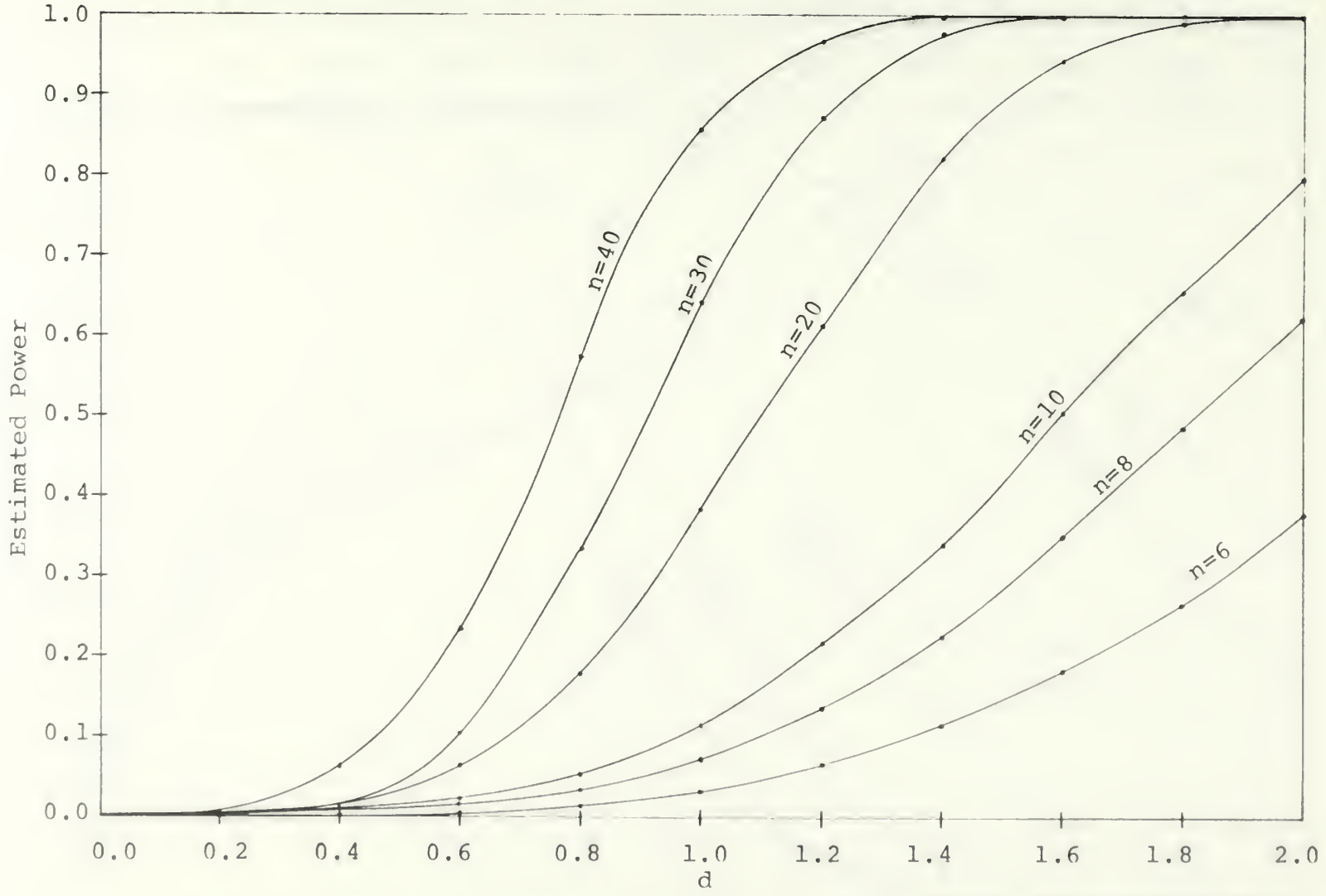

Figure 5. Estimated (unconditioned) power of 1% Scheffé test of four means

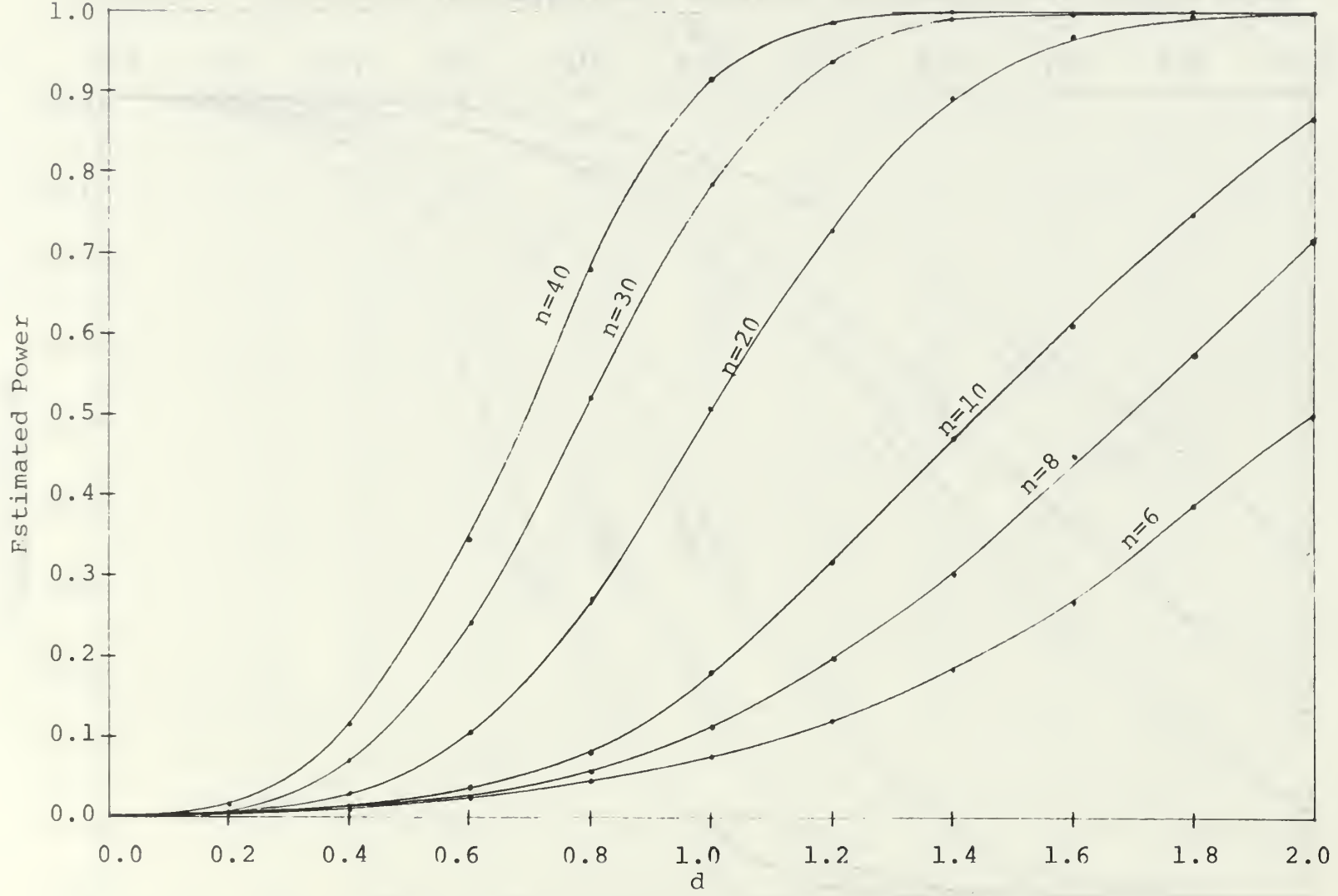

Figure 4. Estimated (unconditioned) power of 1% Scheffé test of three means

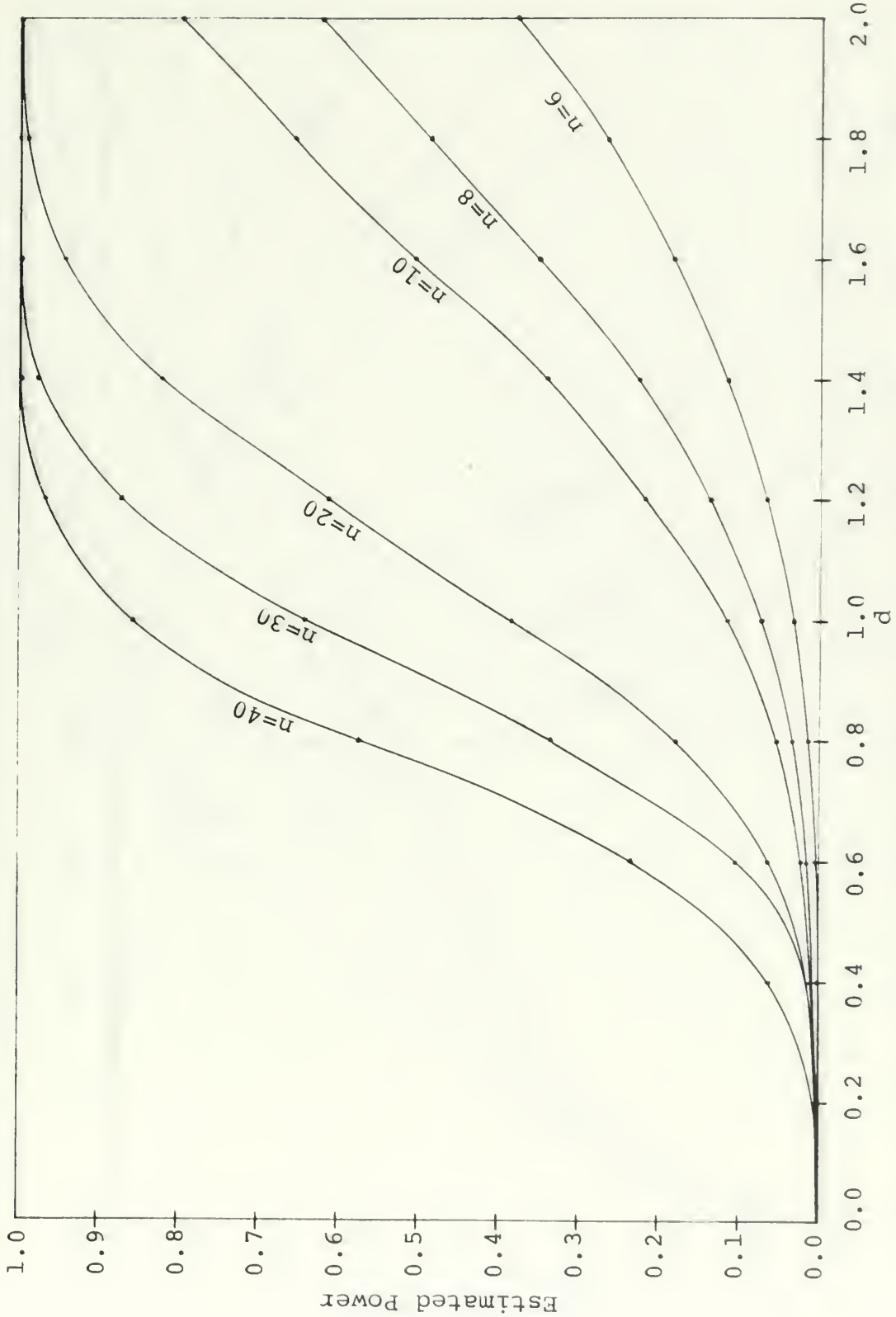

Figure 5. Estimated (unconditioned) power of 1% Scheffé test of four means

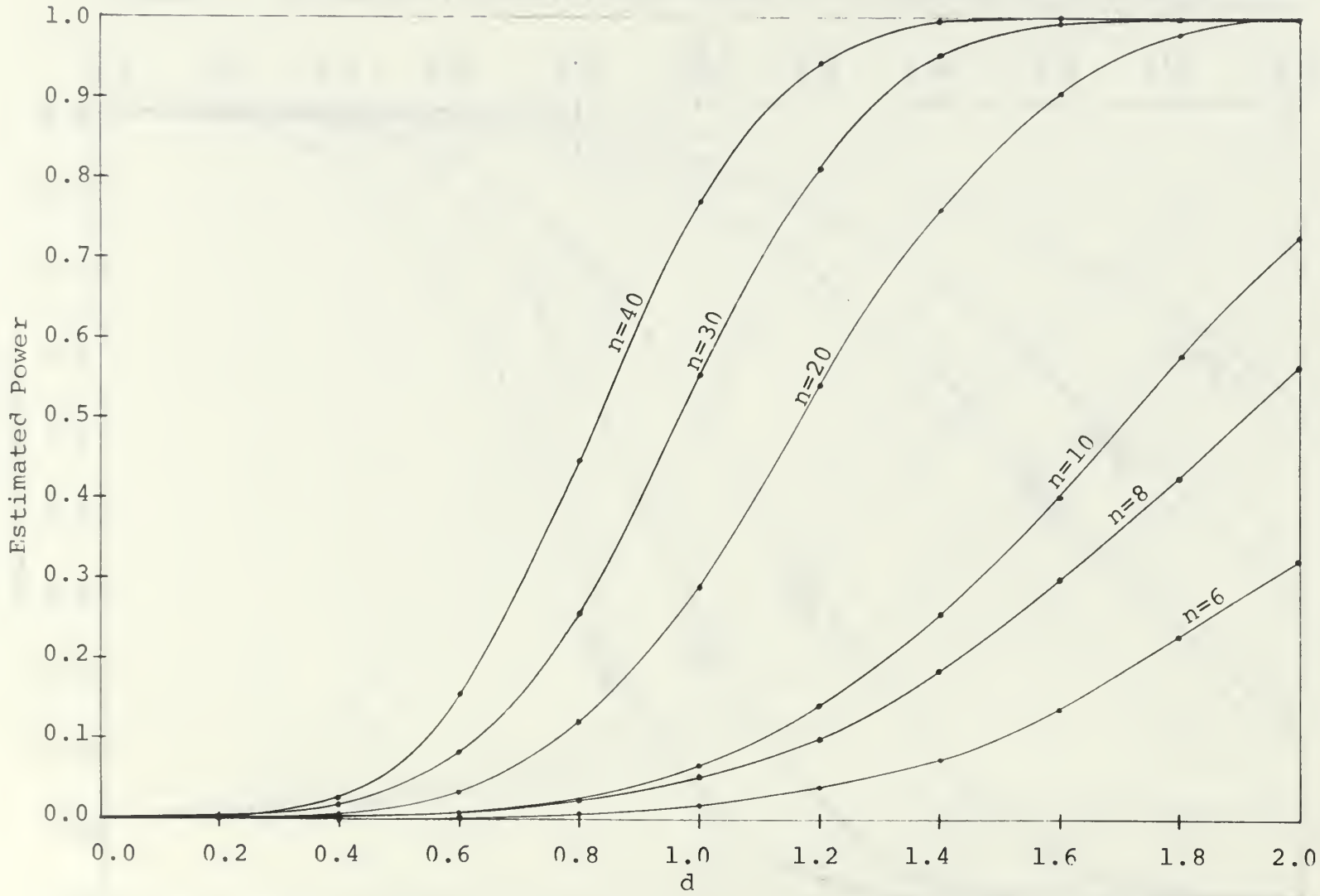

Figure 6. Fstimated (unconditioned) power of 1% Scheffé test of five means

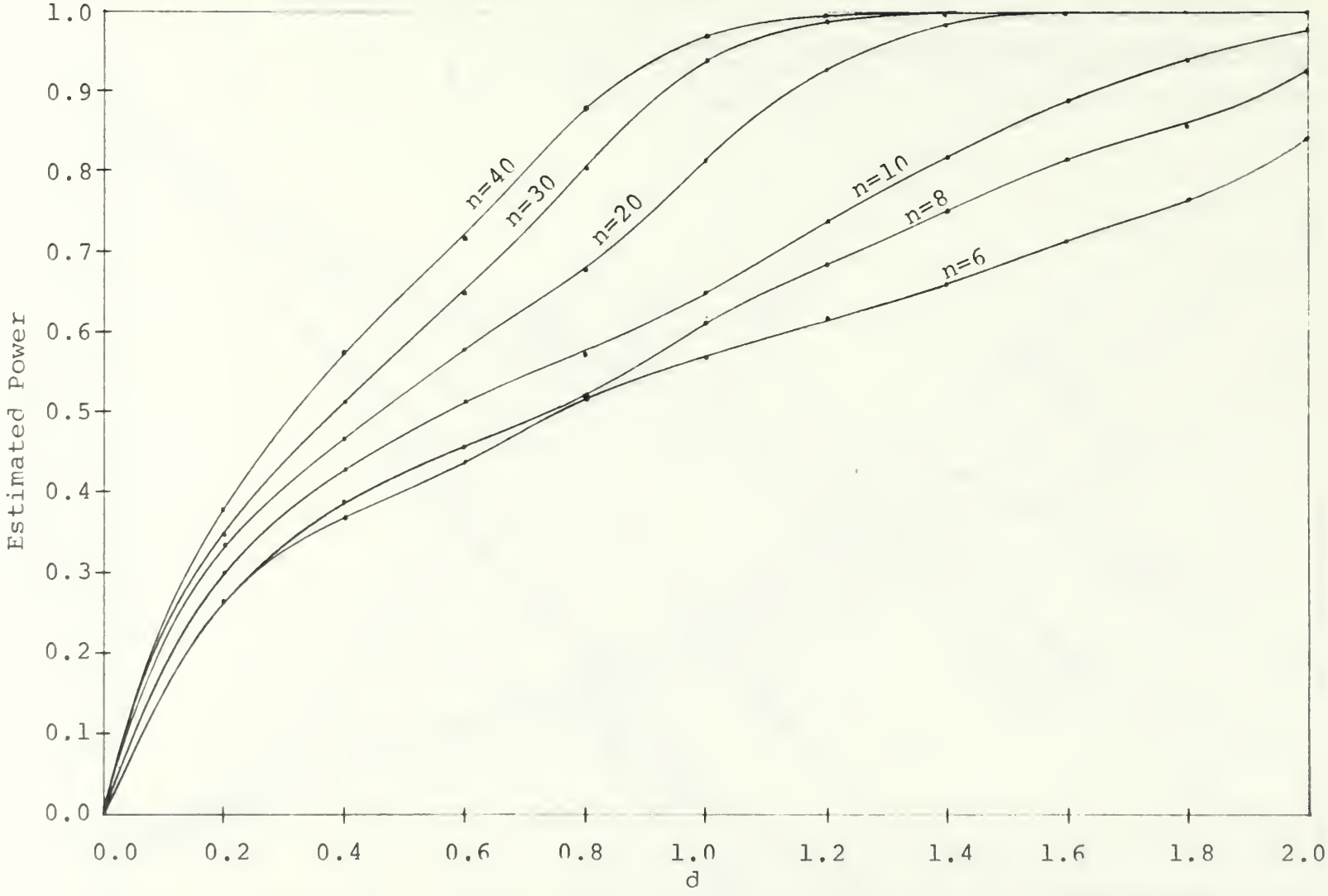

Figure 7. Estimated (conditioned) power of 5% Scheffé test of three means

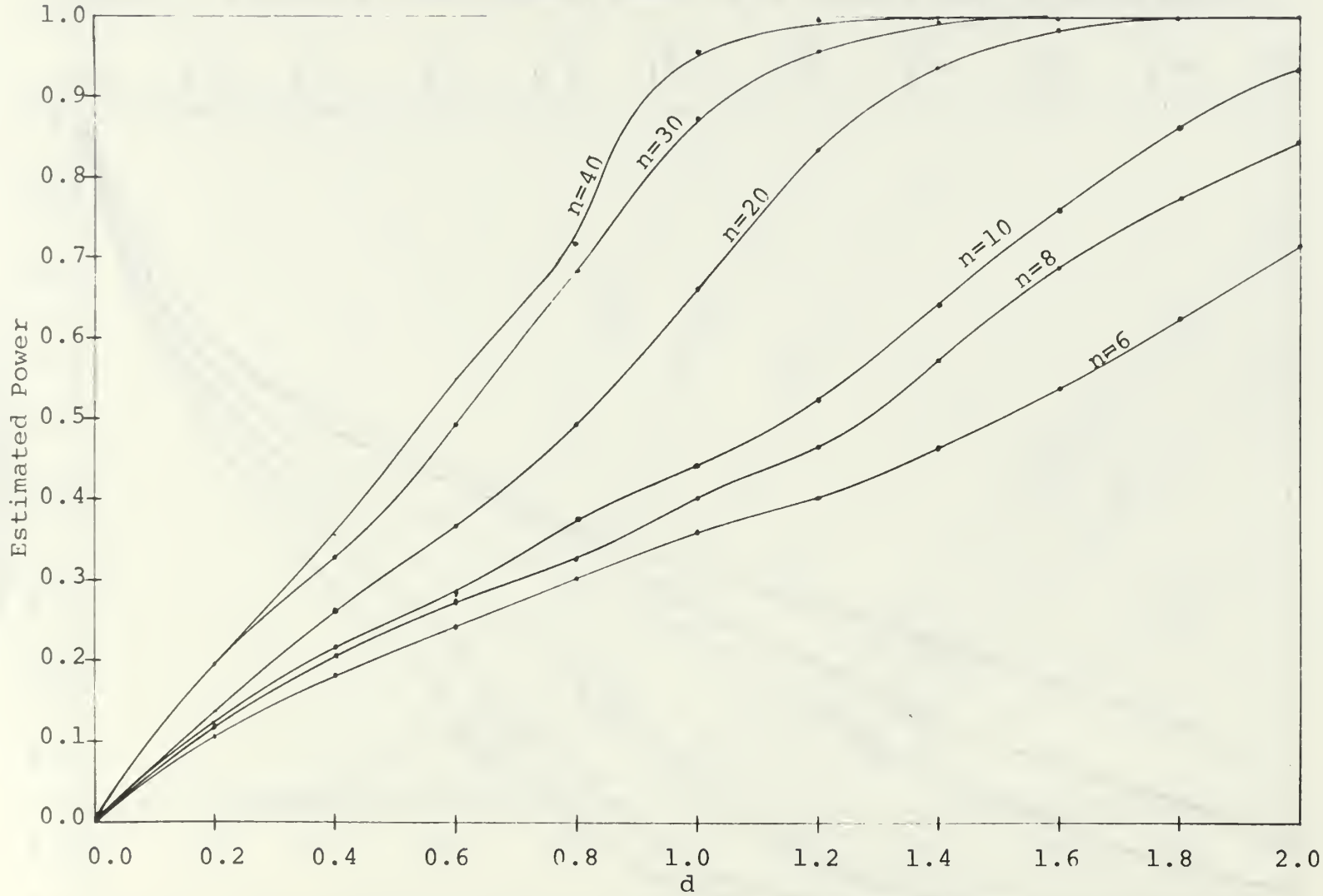

Figure 8. Estimated (conditioned) power of 5% Scheffé test of four means

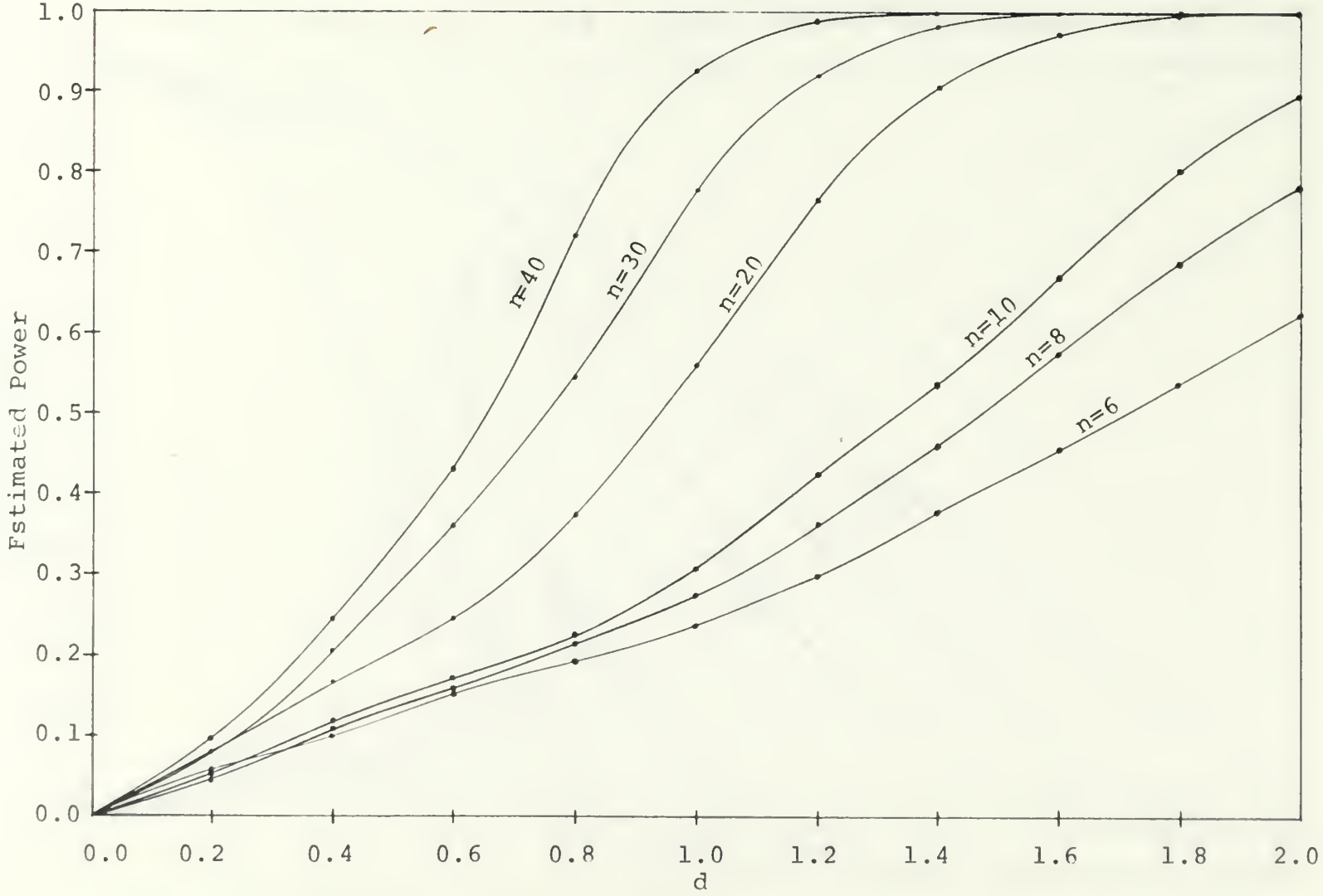

Figure 9. Estimated (conditioned) power of 5% Scheffé test of five means

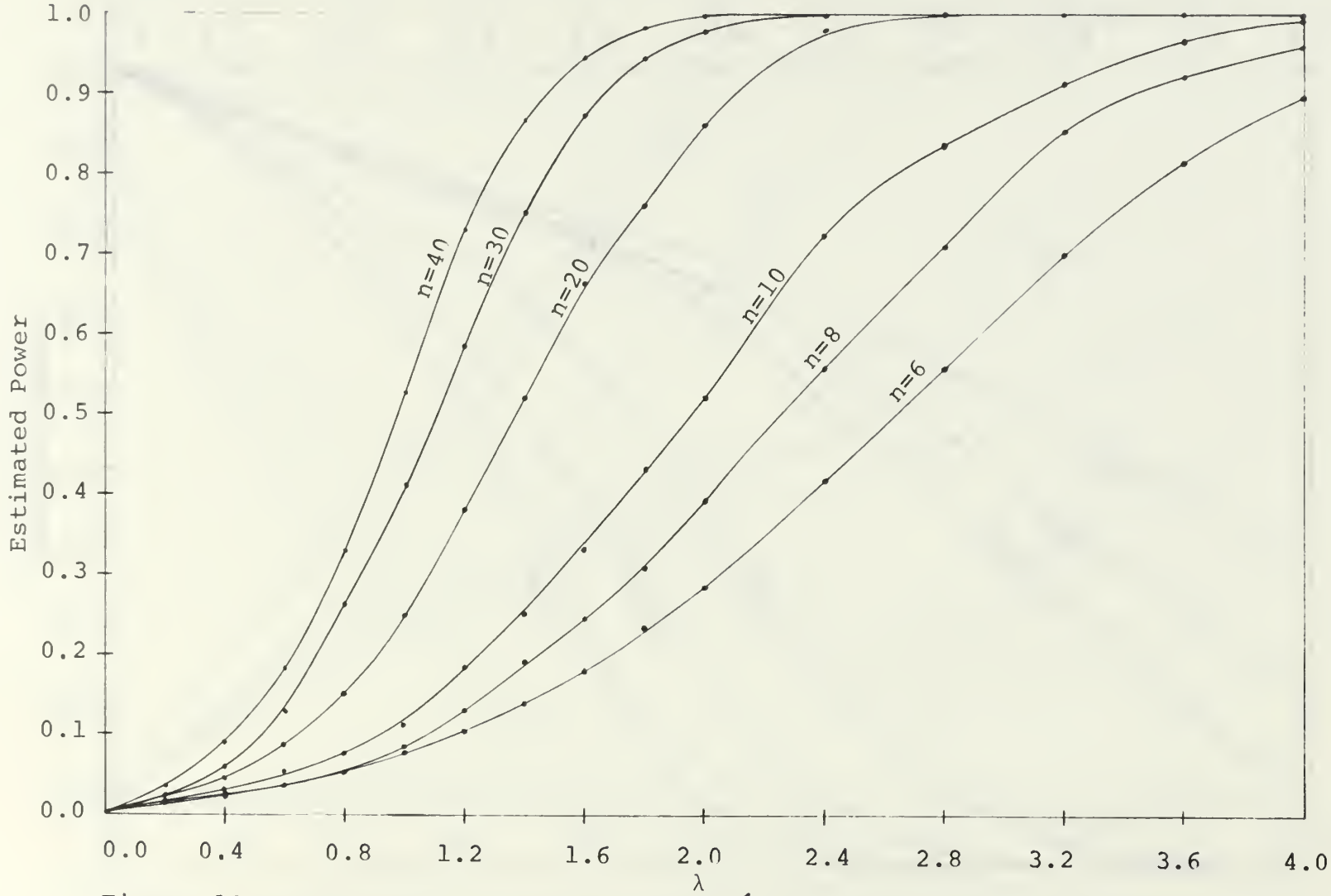

Figure 10. Estimated power of 5% Scheffé test of a linear combination of 3 means

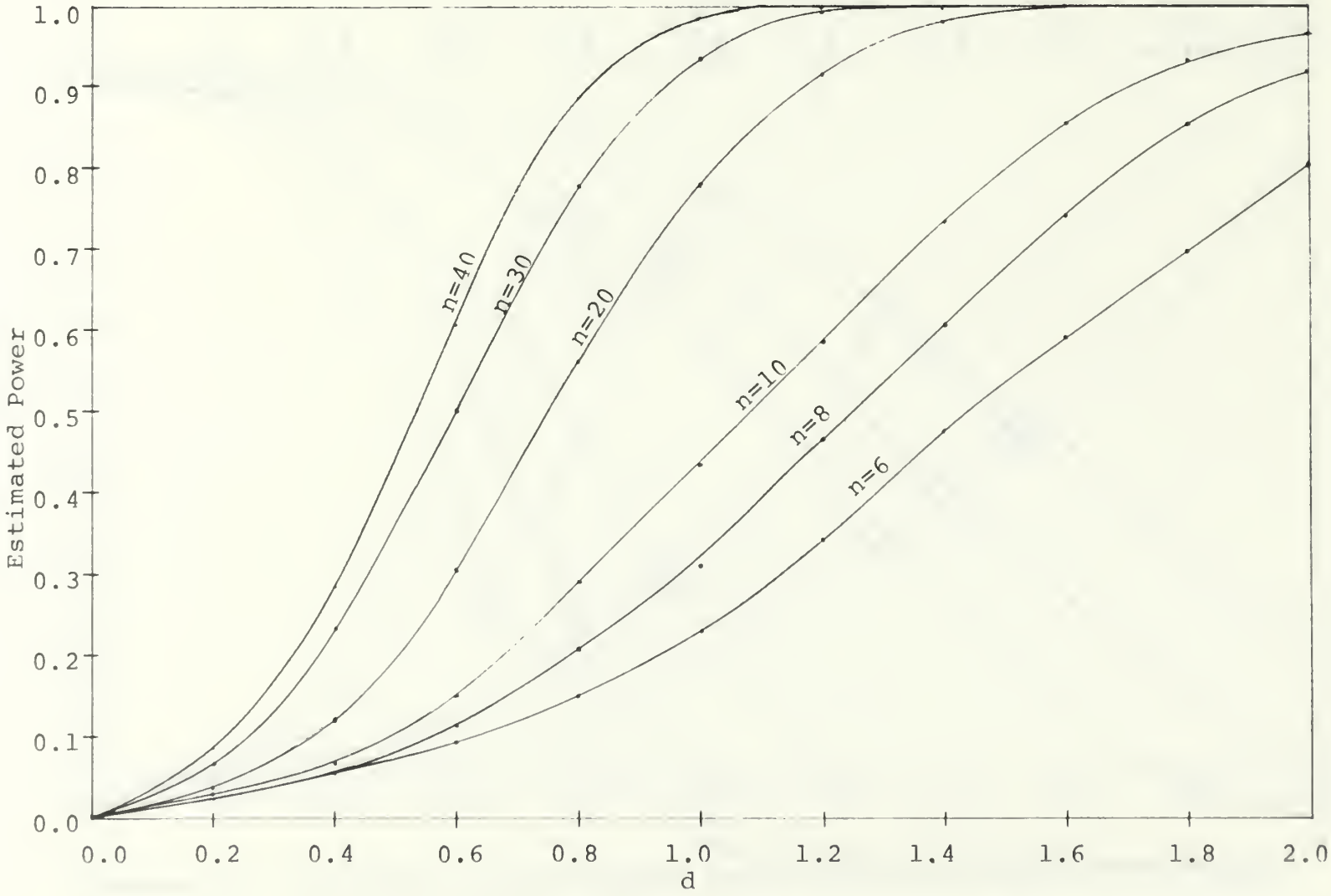

Figure 11. Estimated (unconditioned) power of 5% Tukey test of three means

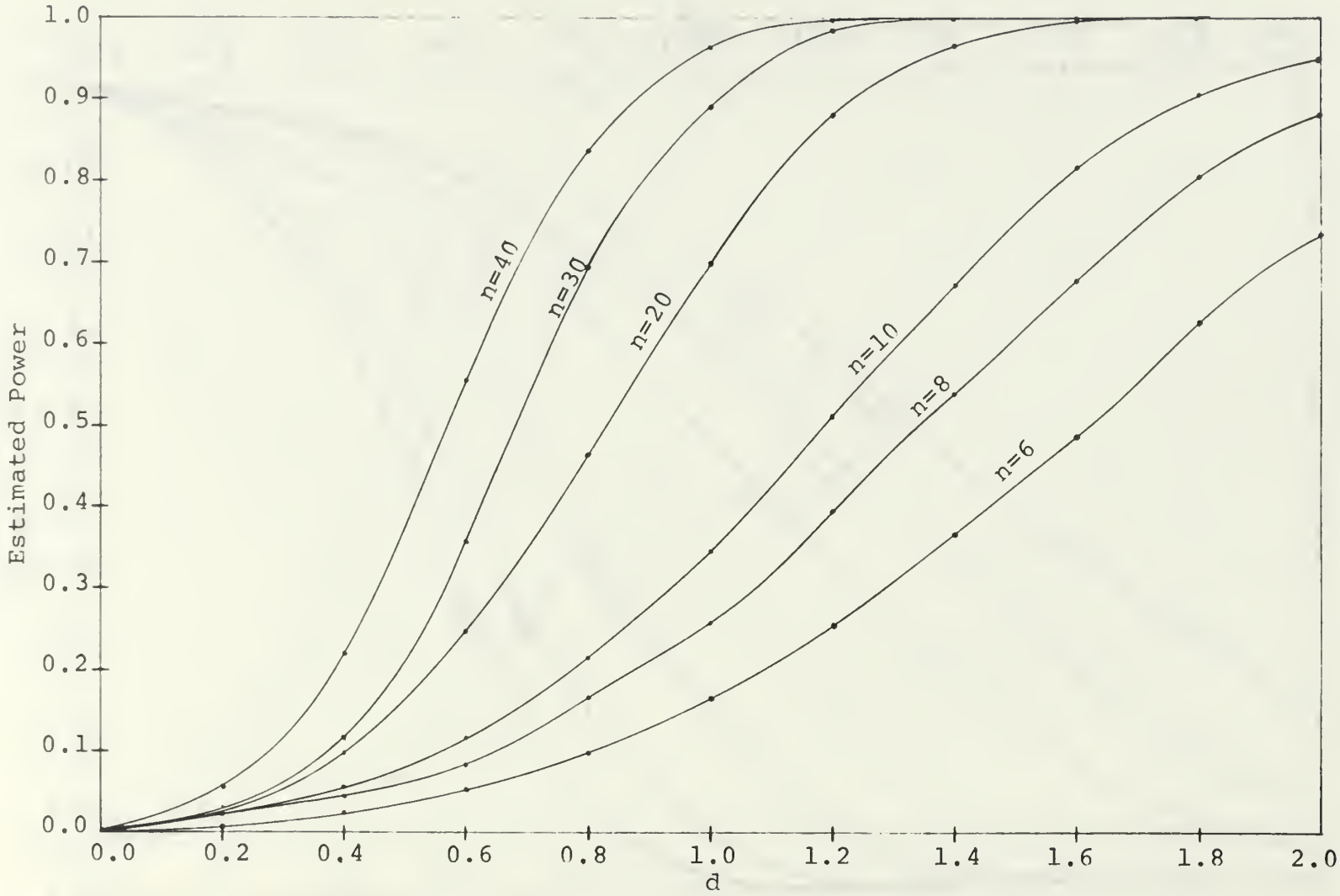

Figure 12. Estimated (unconditioned) power of 5% Tukey test of four means

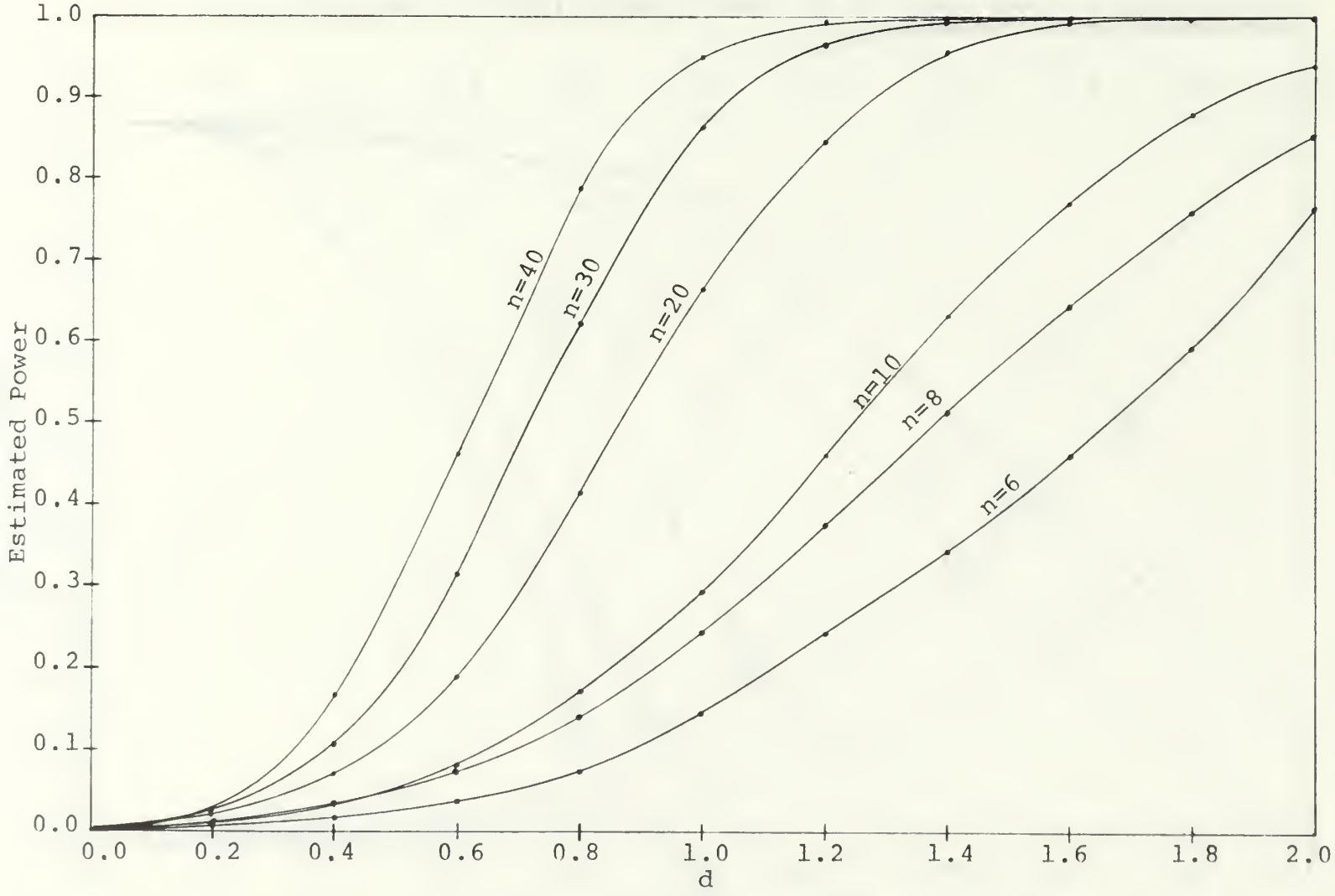

Figure 13. Estimated (unconditioned) power of 5% Tukey test of five means

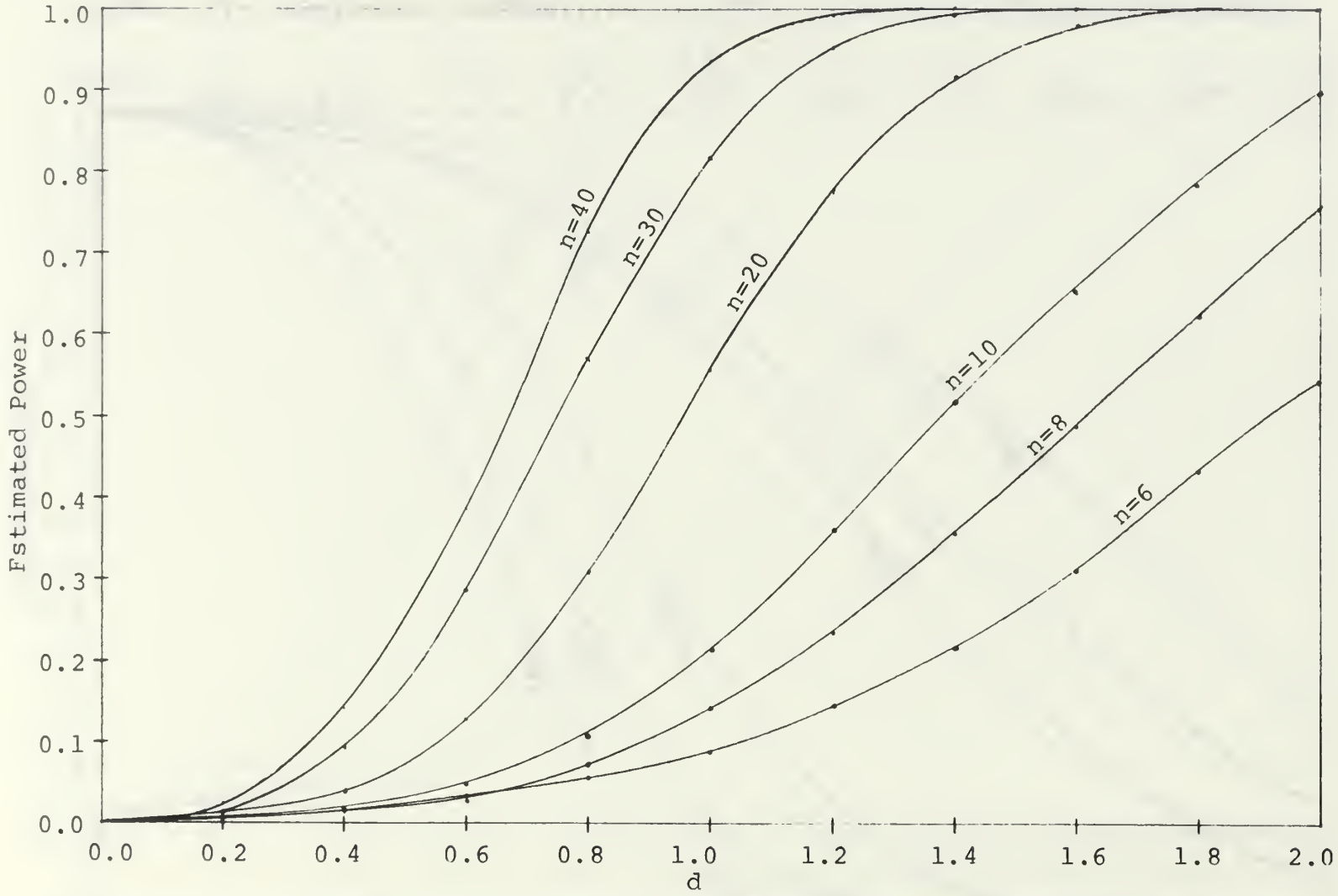

Figure 14. Estimated (unconditioned) power of 1% Tukey test of three means

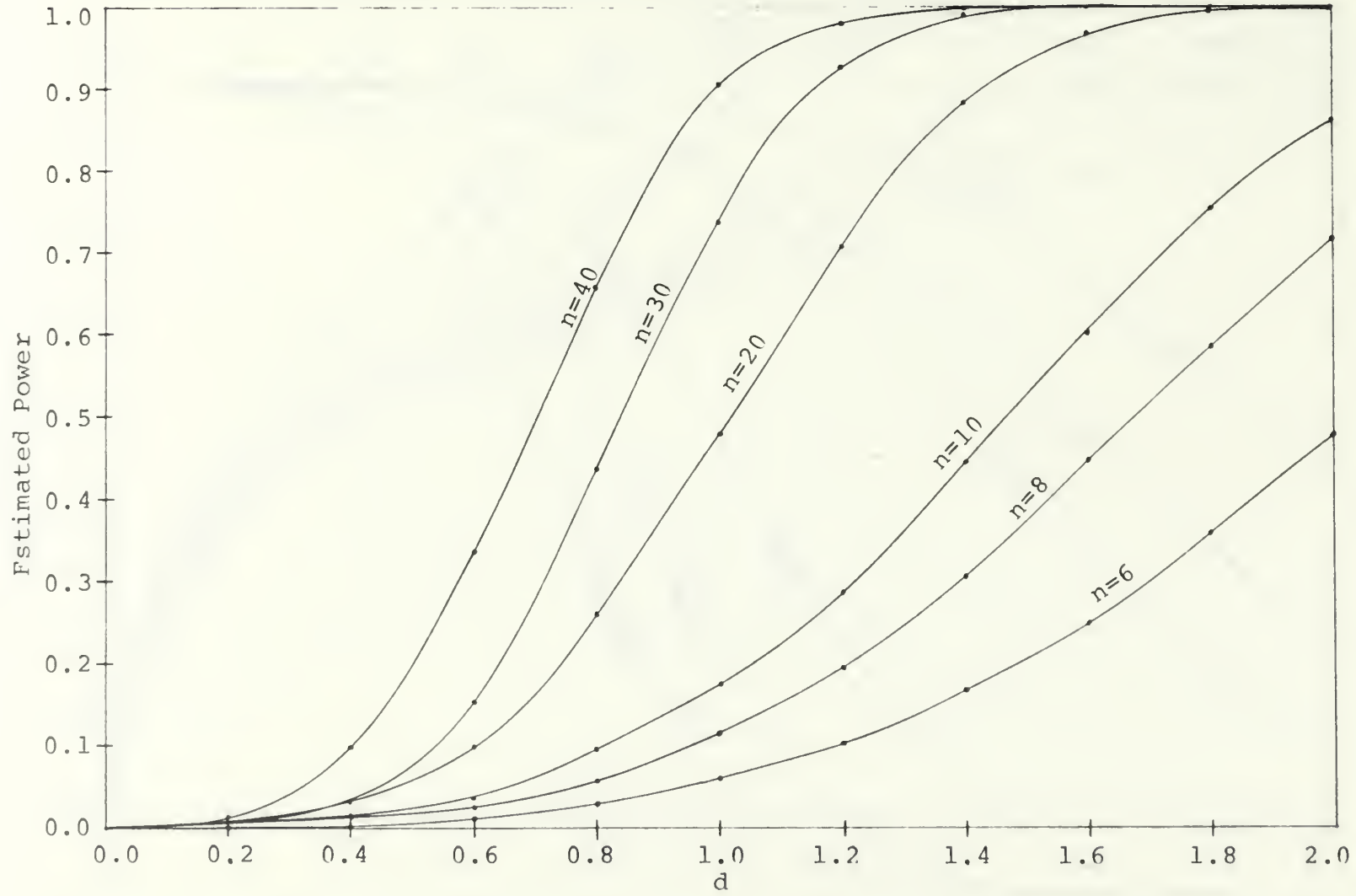

Figure 15. Fstimated (unconditioned) power of 1% Tukey test of four means

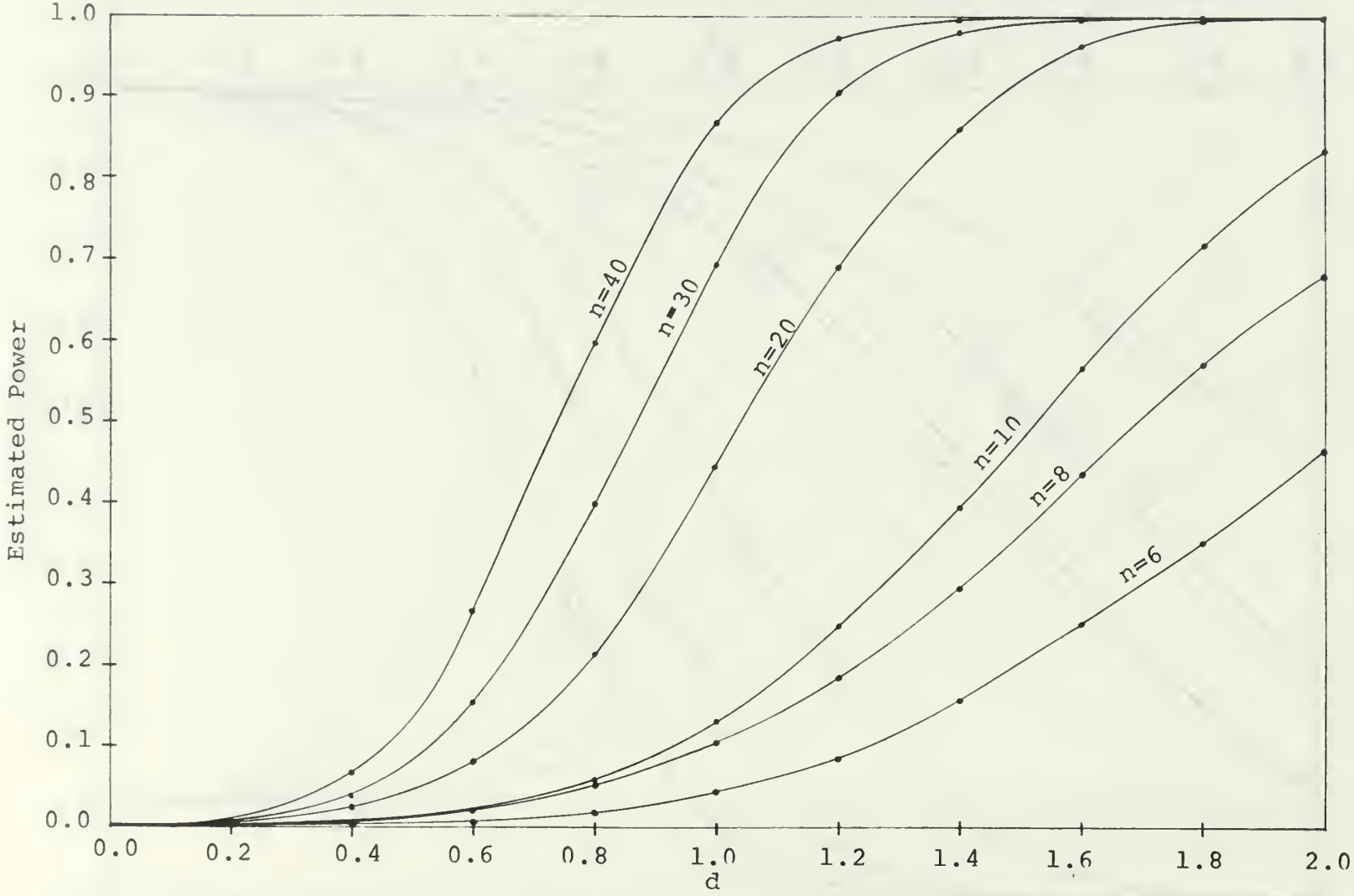

Figure 16. Estimated (unconditioned) power of 1% Tukey test of five means

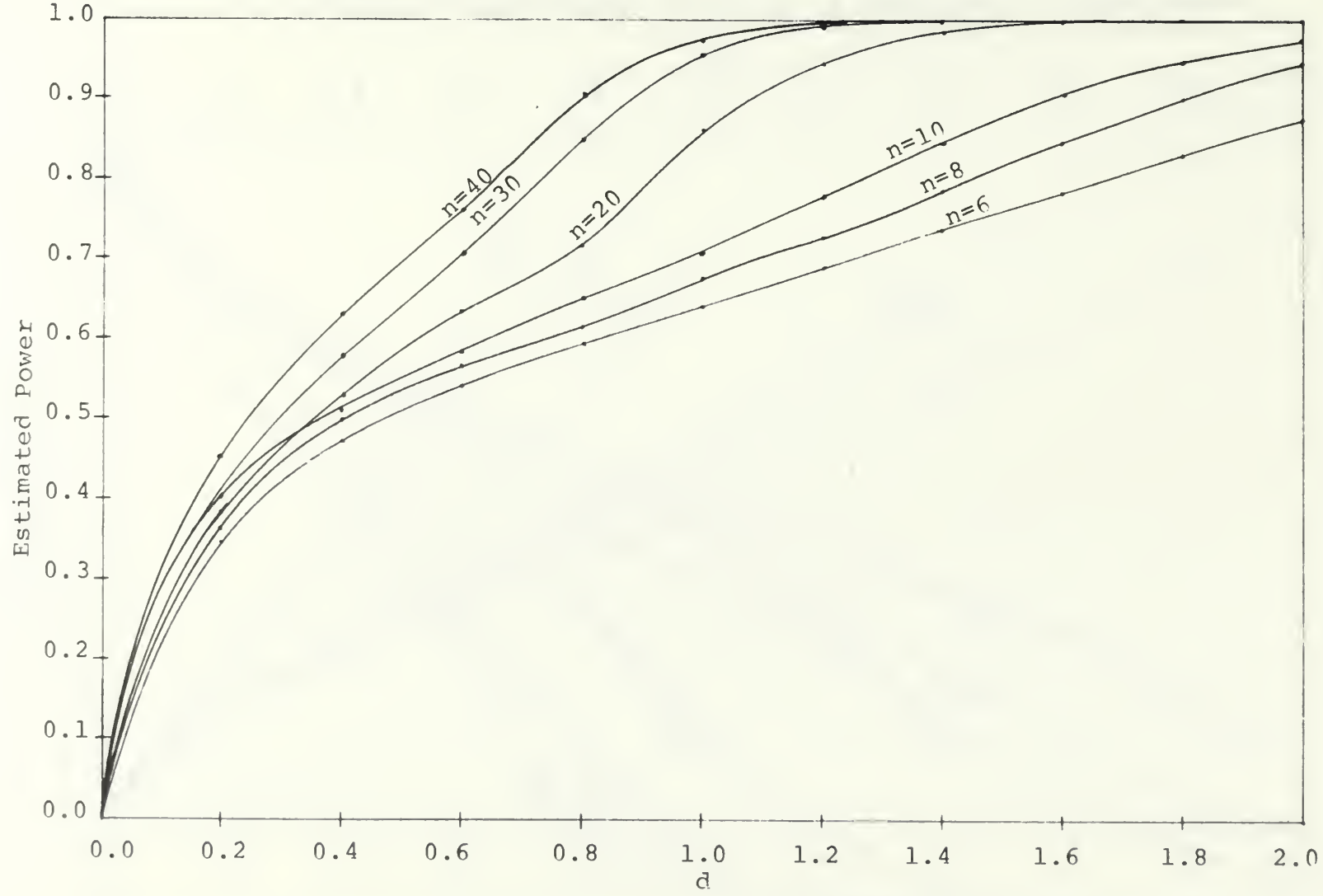

Figure 17. Estimated (conditioned) power of 5% Tukey test of three means

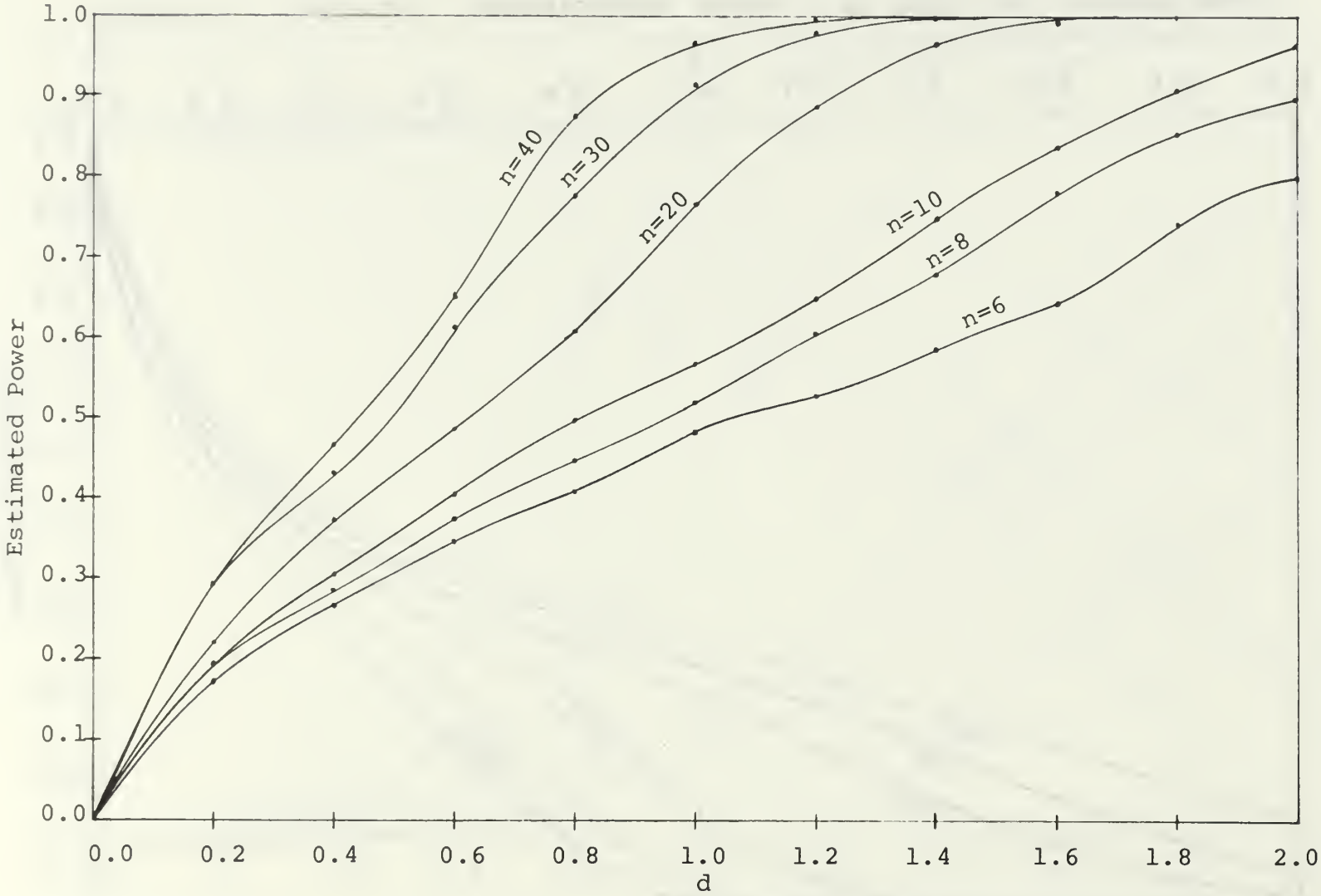

Figure 18. Estimated (conditioned) power of 5% Tukey test of four means

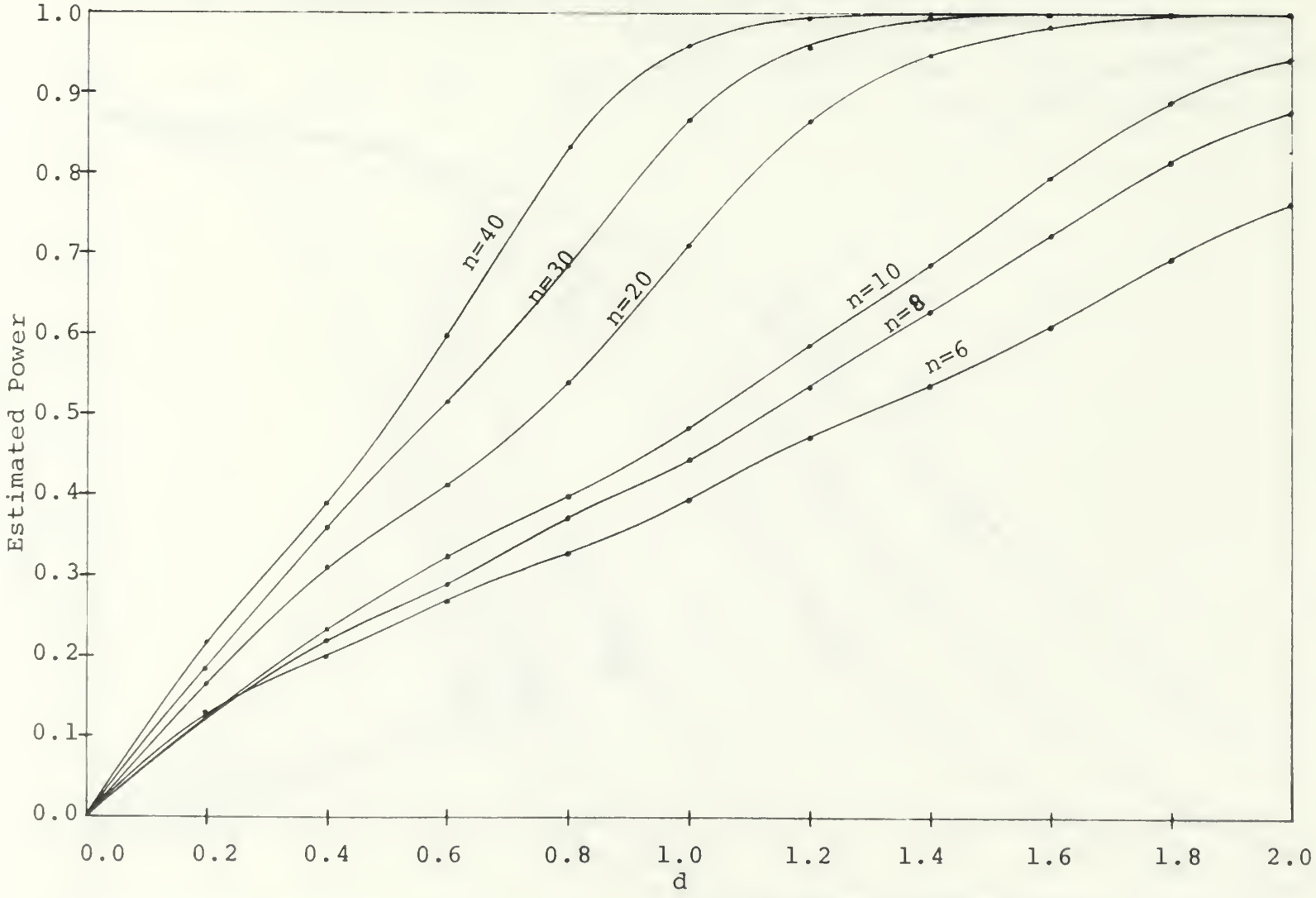

Figure 19. Estimated (conditioned) power of 5% Tukey test of five means

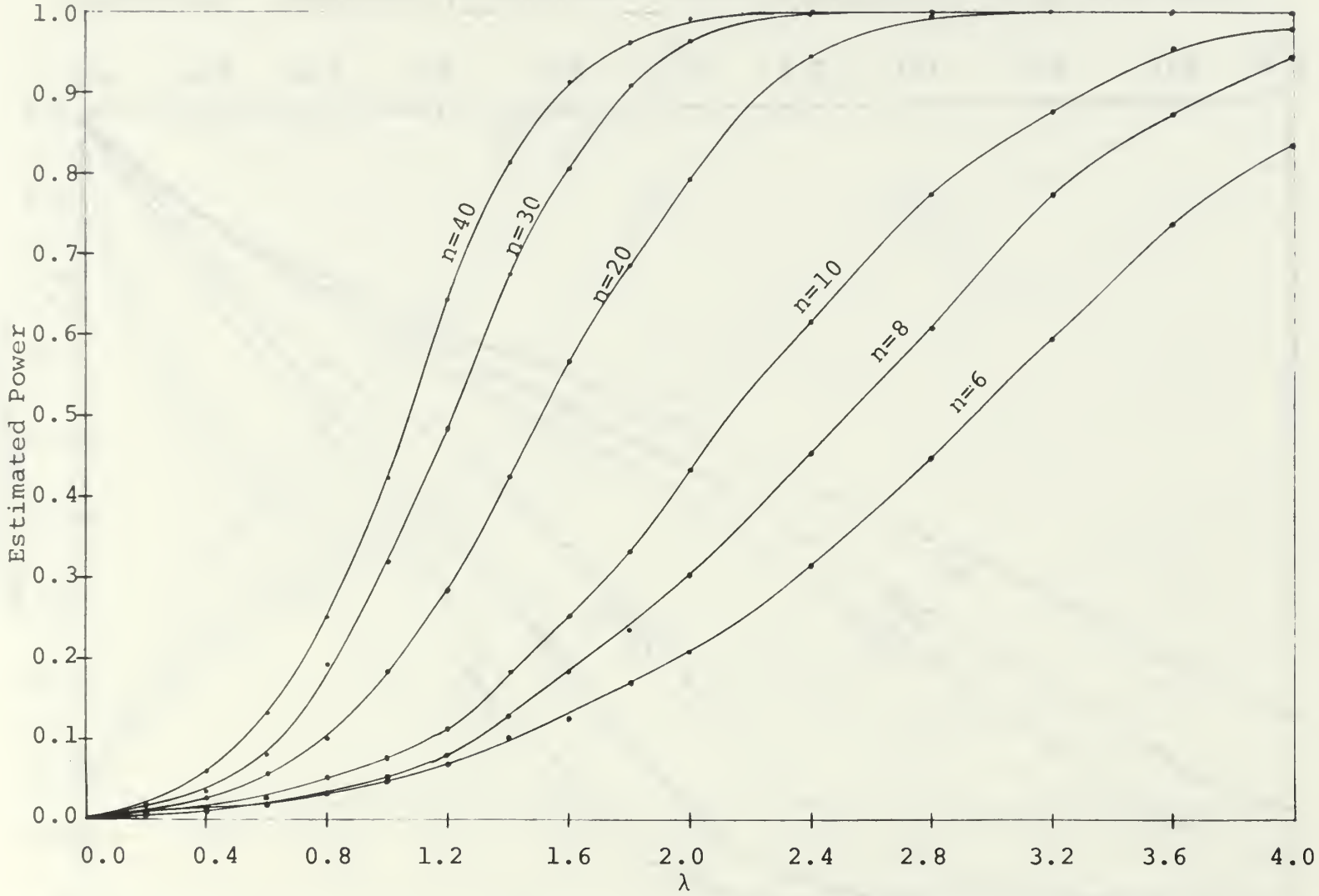

Figure 20. Estimated power of 5% Tukey test of a linear combination of three means

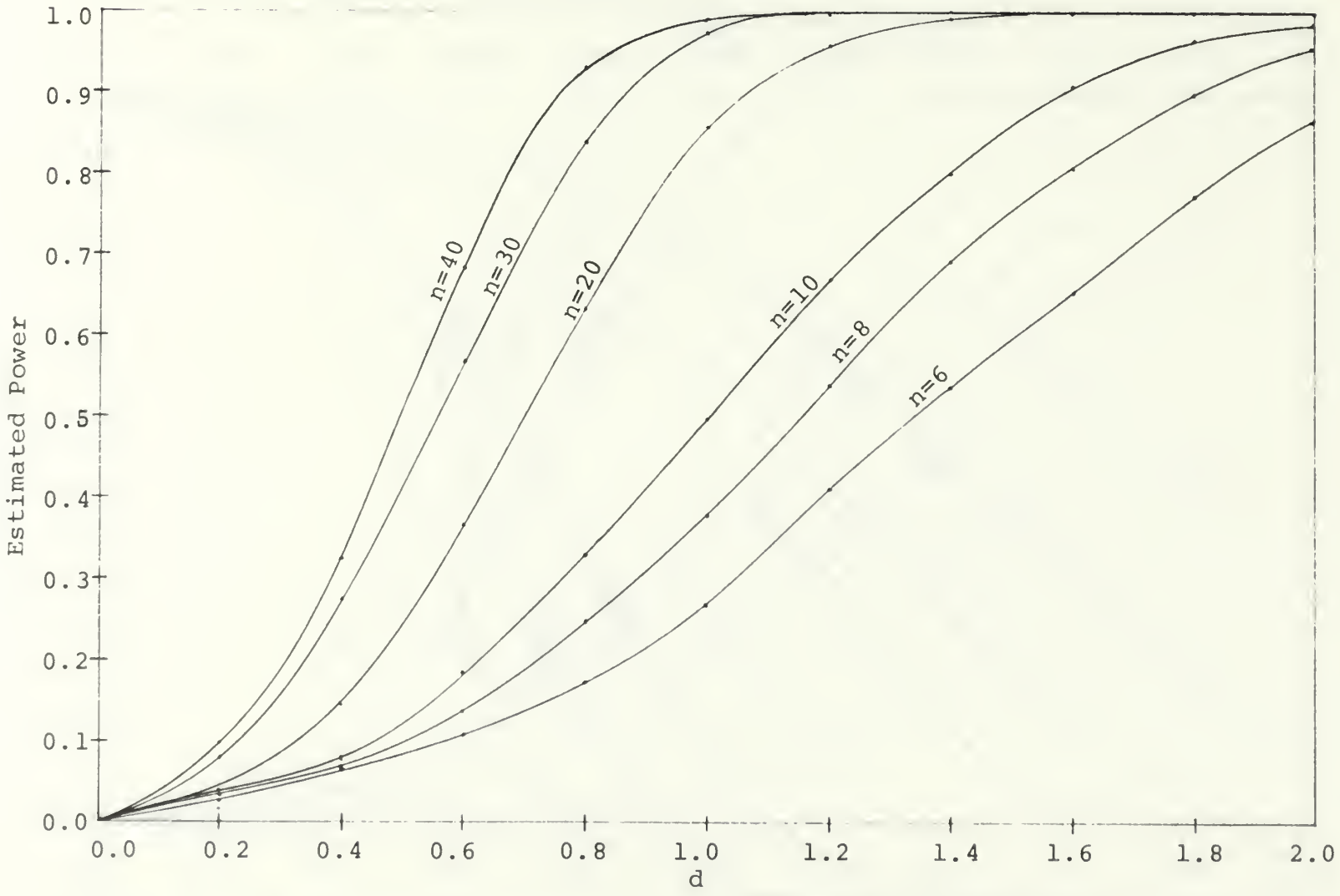

Figure 21. Estimated (unconditioned) power of 5% S-N-K test of three means

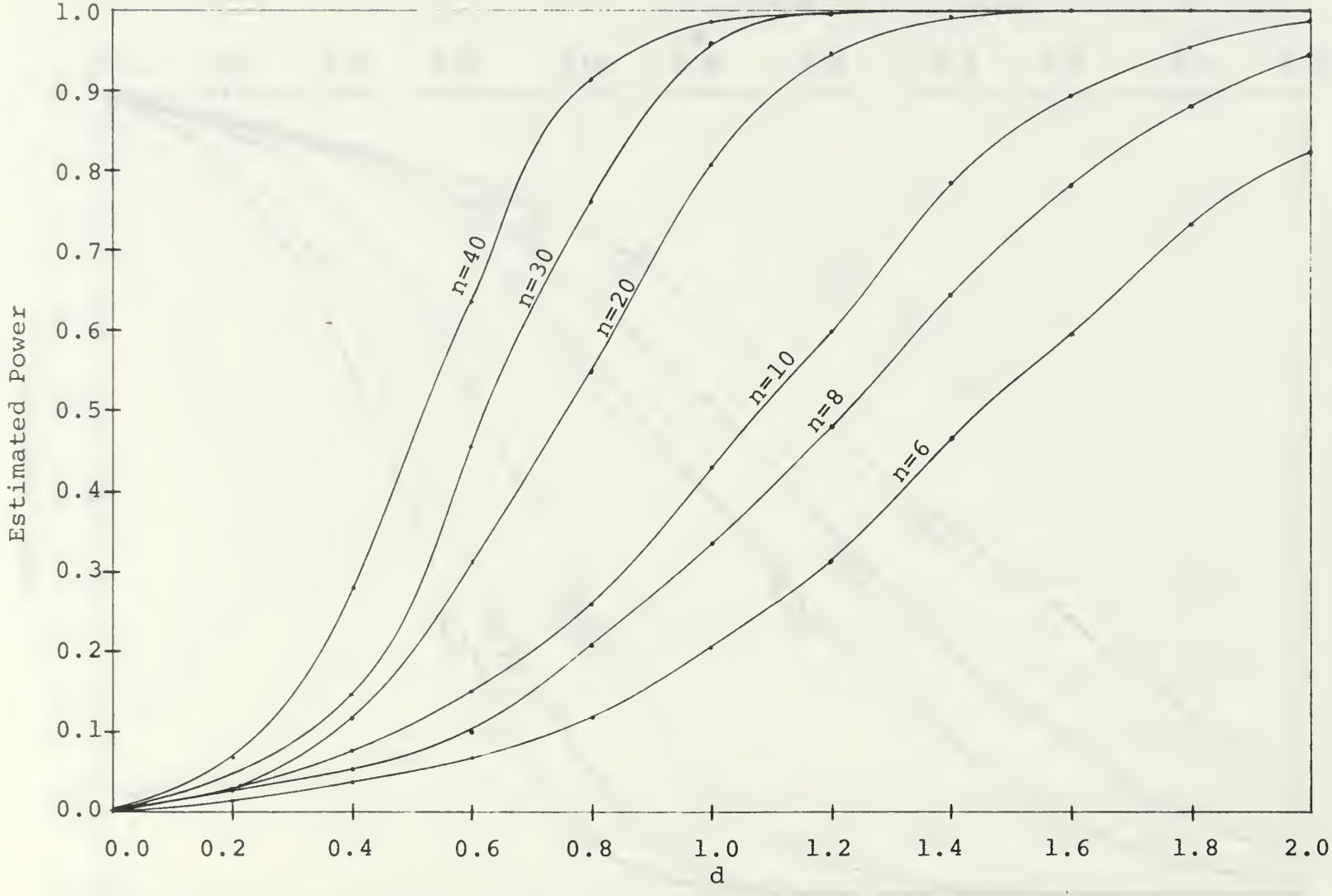

Figure 22. Estimated (unconditioned) power of 5% S-N-K test of four means

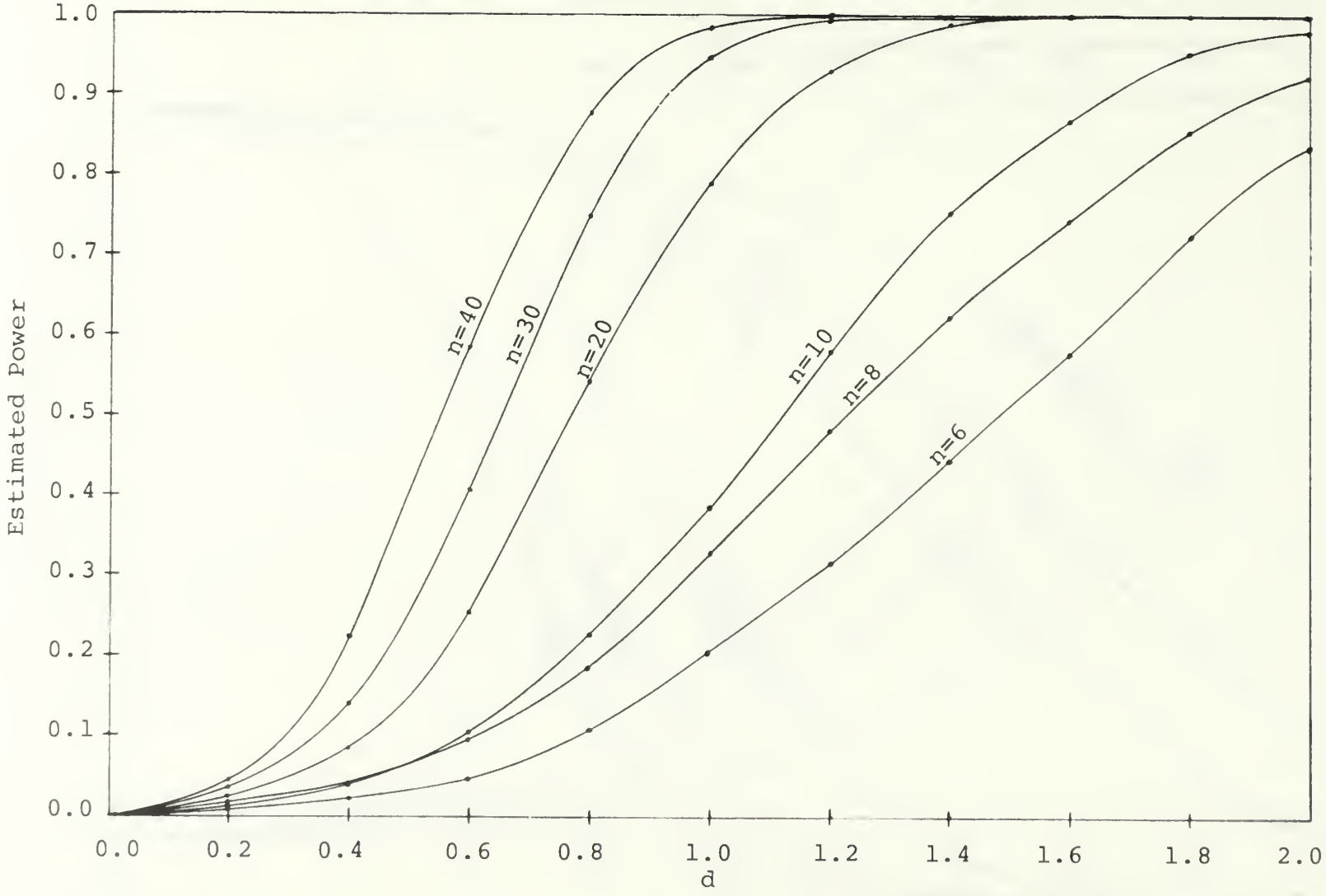

Figure 23. Estimated (unconditioned) power of 5% S-N-K test of five means

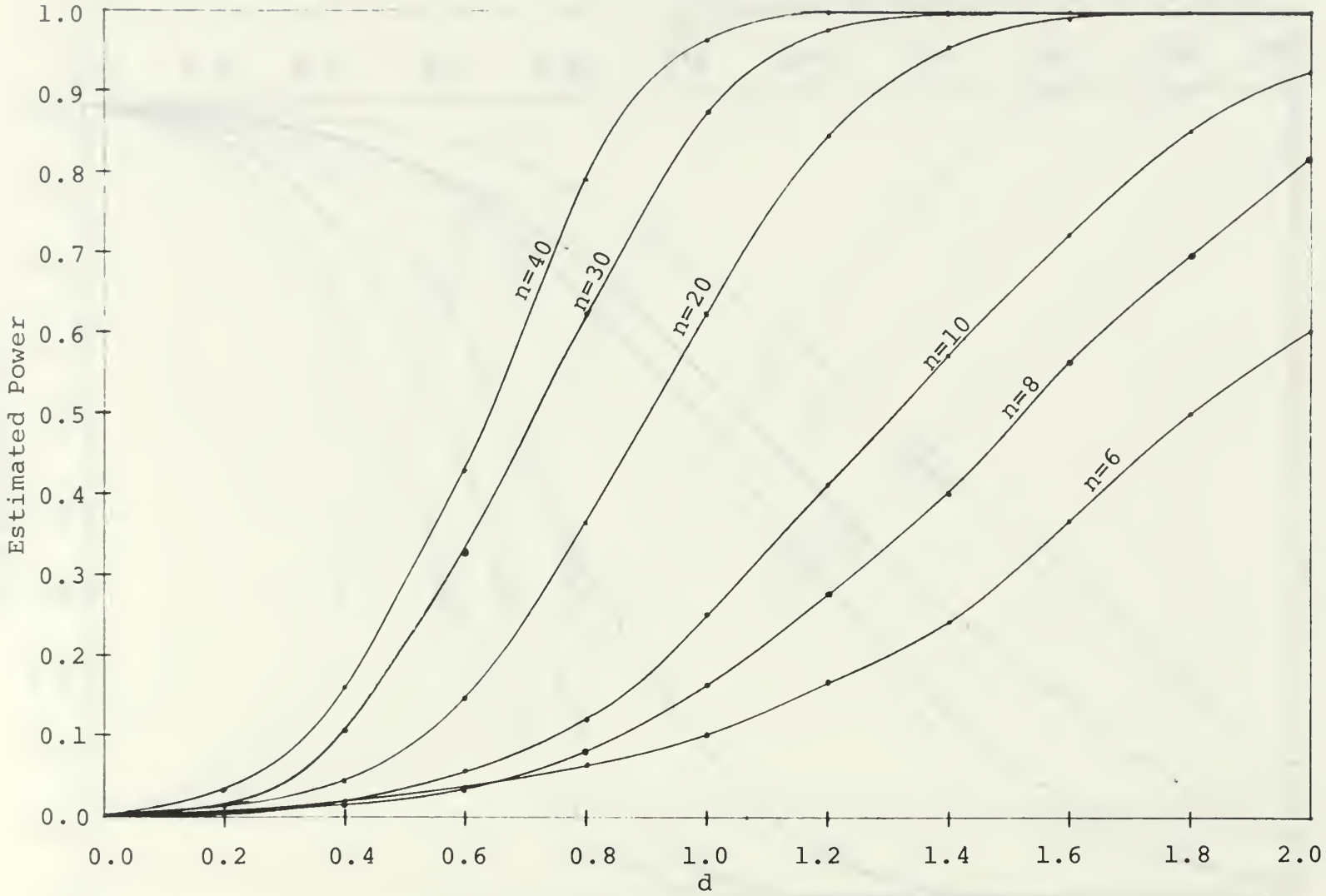

Figure 24. Estimated (unconditioned) power of 1% S-N-K test of three means

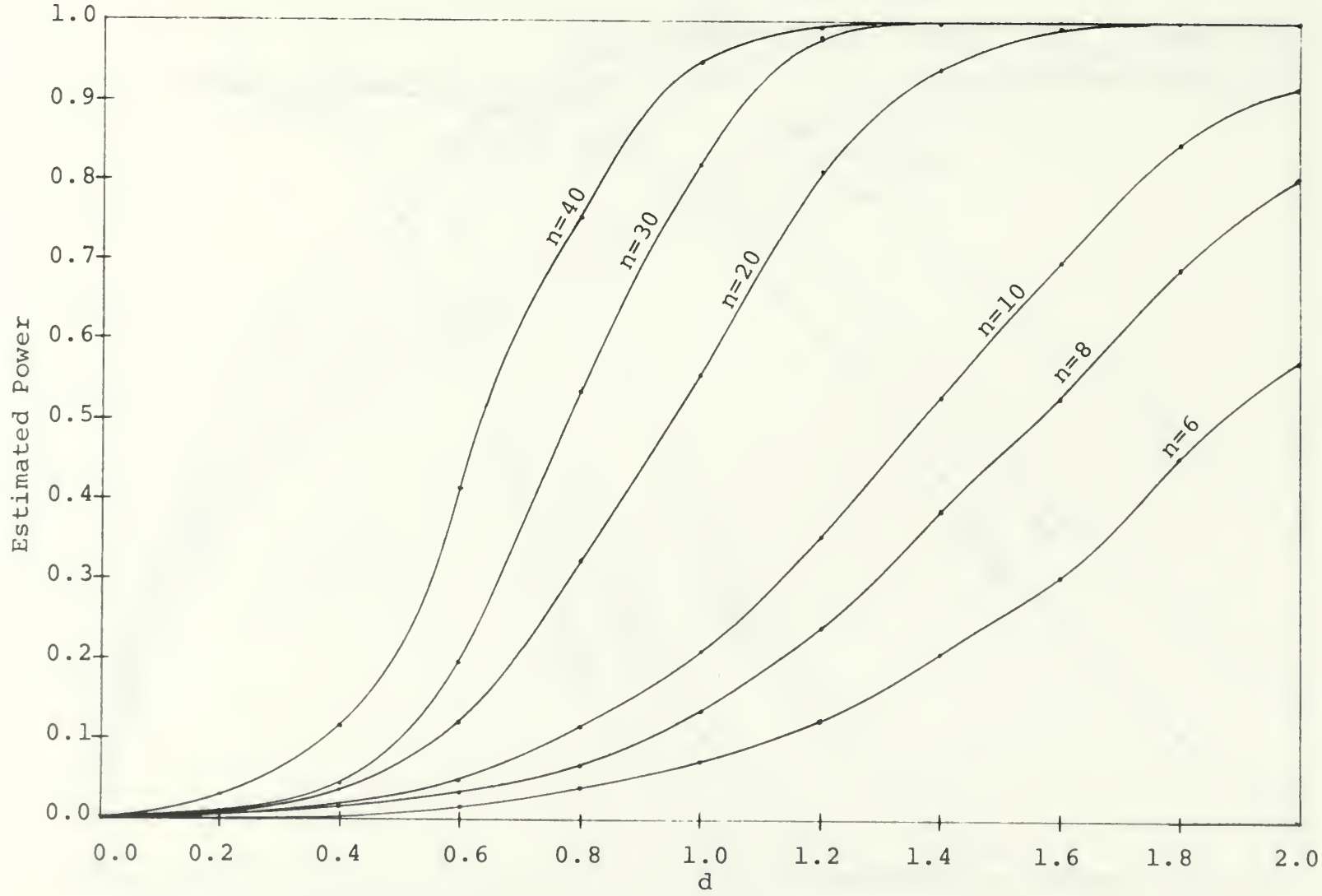

Figure 25. Estimated (unconditioned) power of 1% S-N-K test of four means

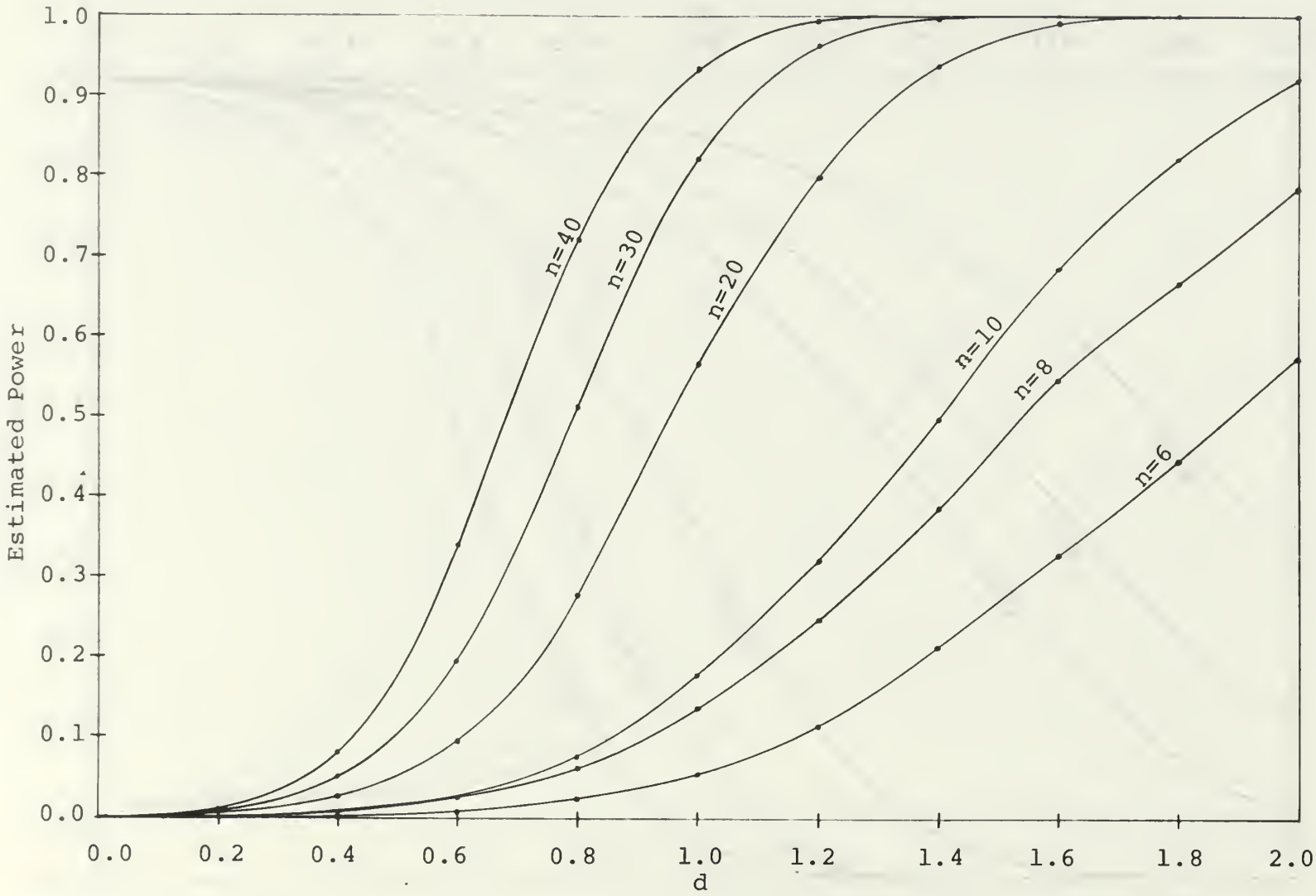

Figure 26. Estimated (unconditioned) power of 1% S-N-K test of five means

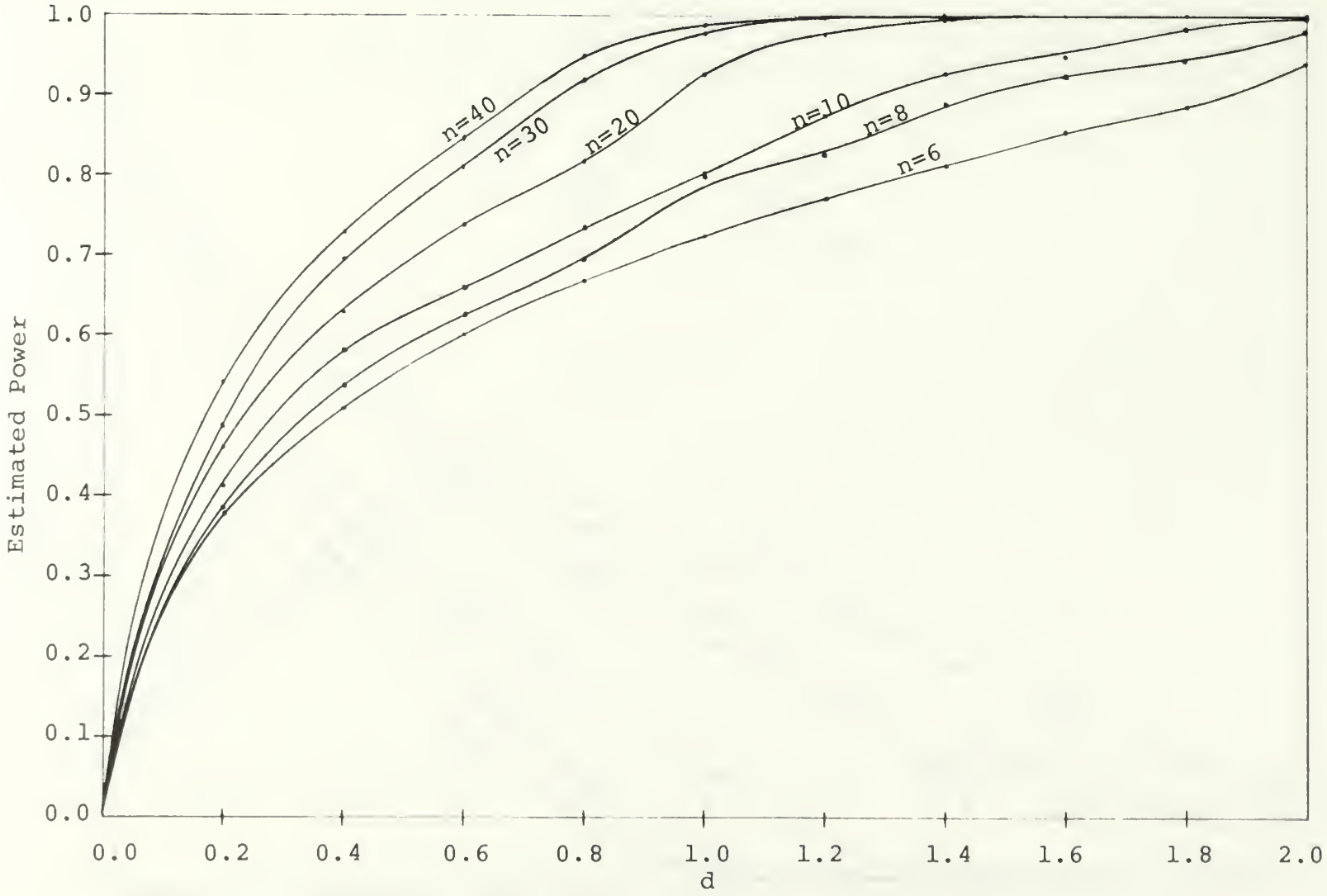

Figure 27. Estimated (conditioned) power of 5% S-N-K test of three means

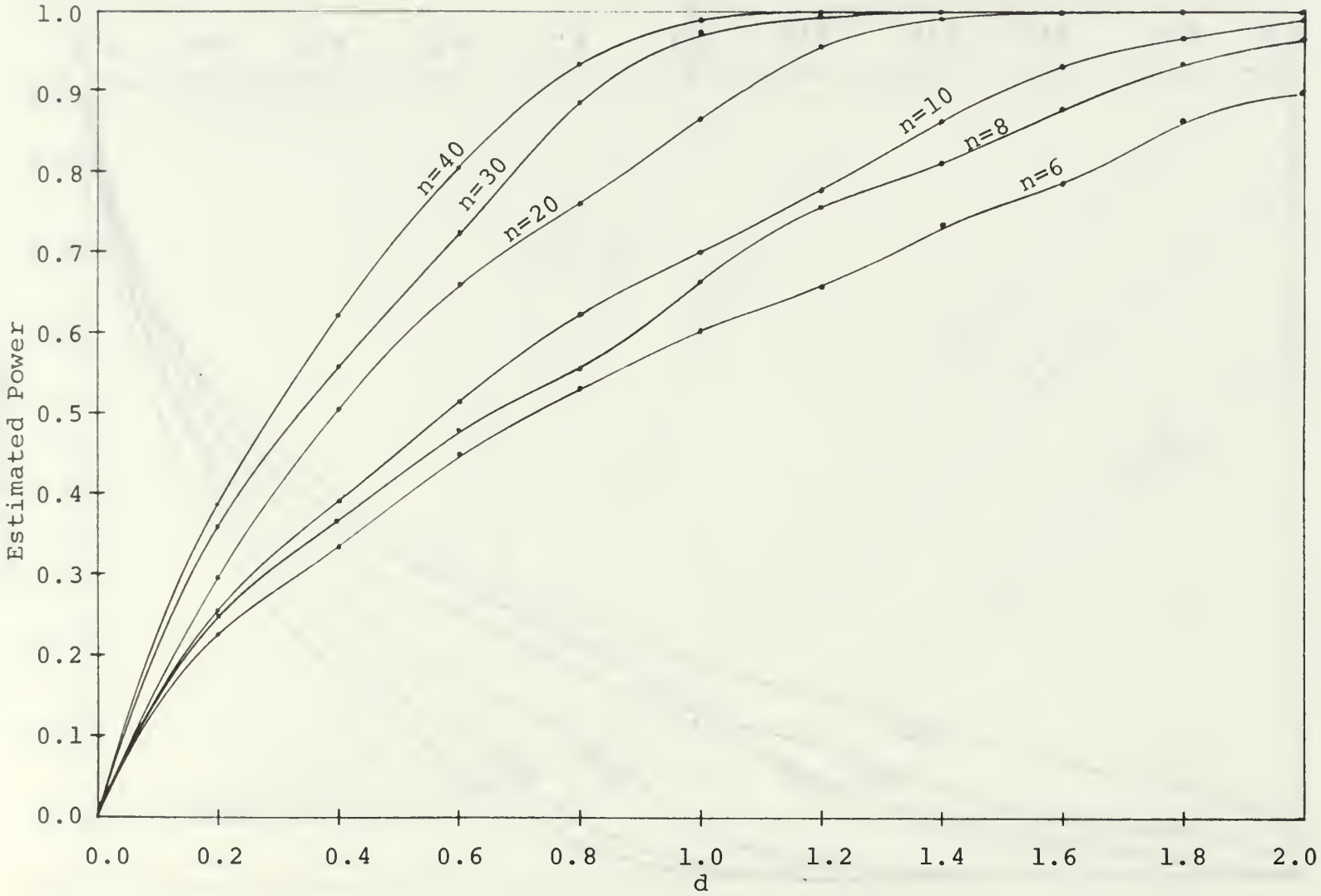

Figure 28. Estimated (conditioned) power of 5% S-N-K test of four means

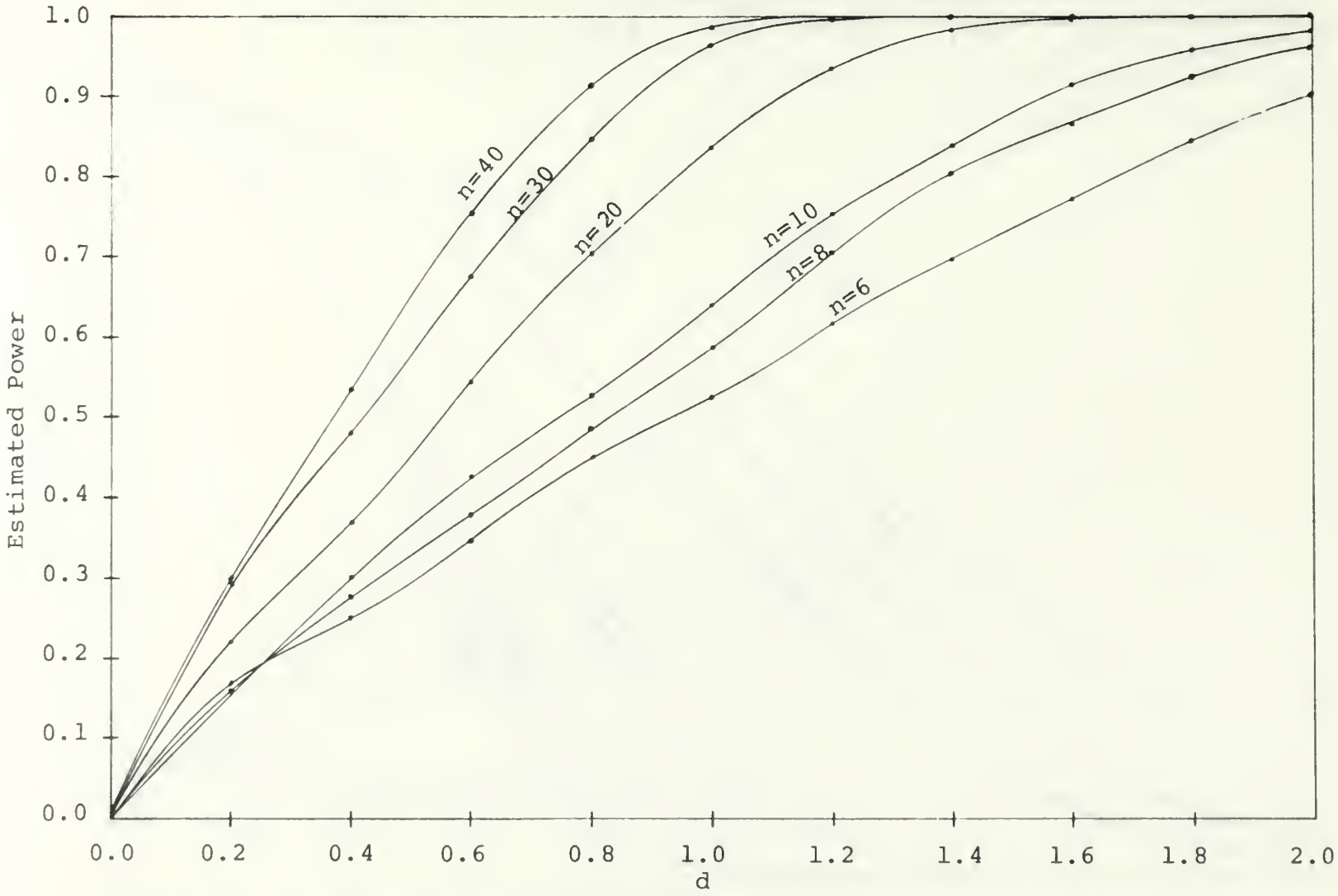

Figure 29. Estimated (conditioned) power of 5% S-N-K test of five means

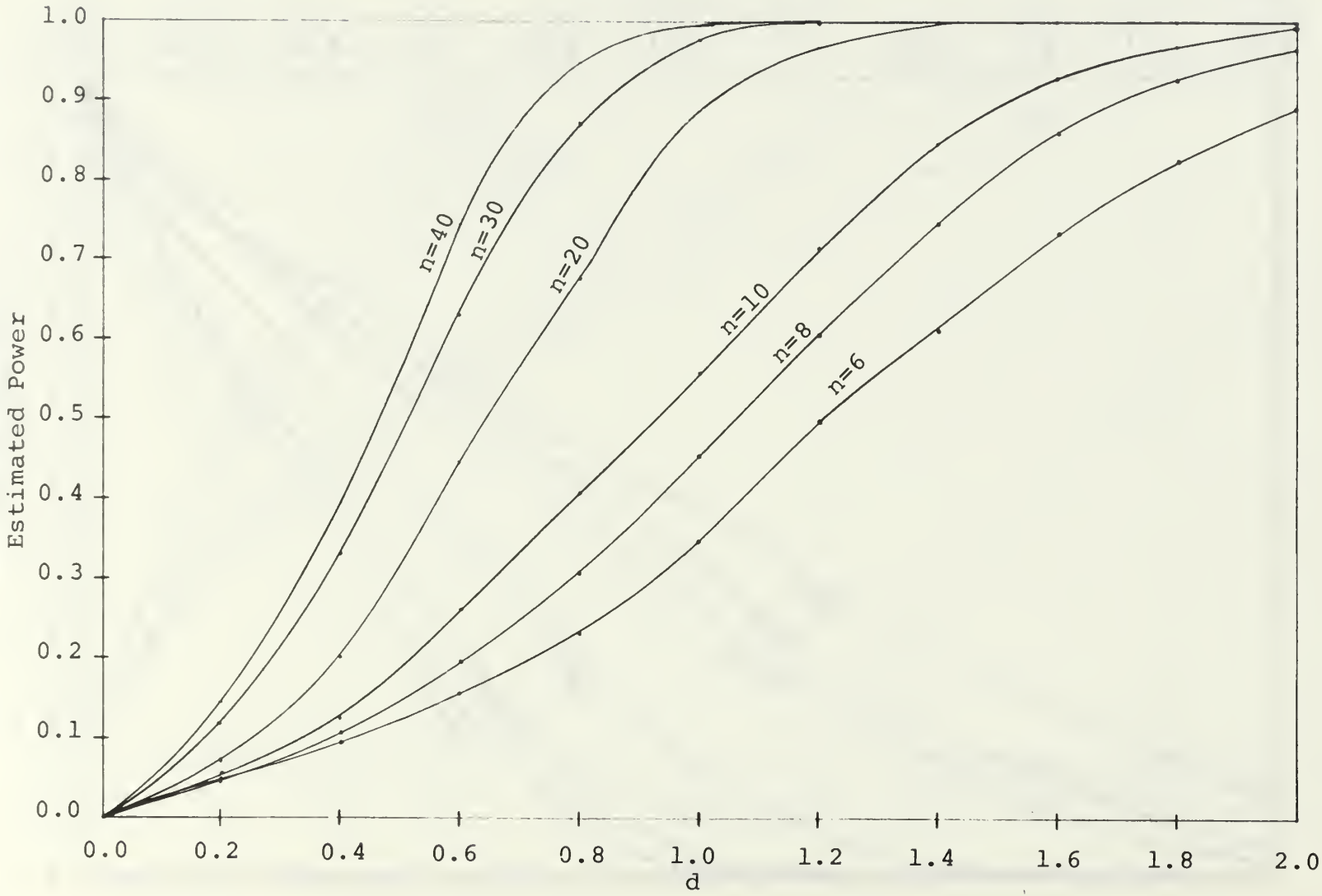

Figure 30. Estimated (unconditioned) power of 5% Duncan test of three means

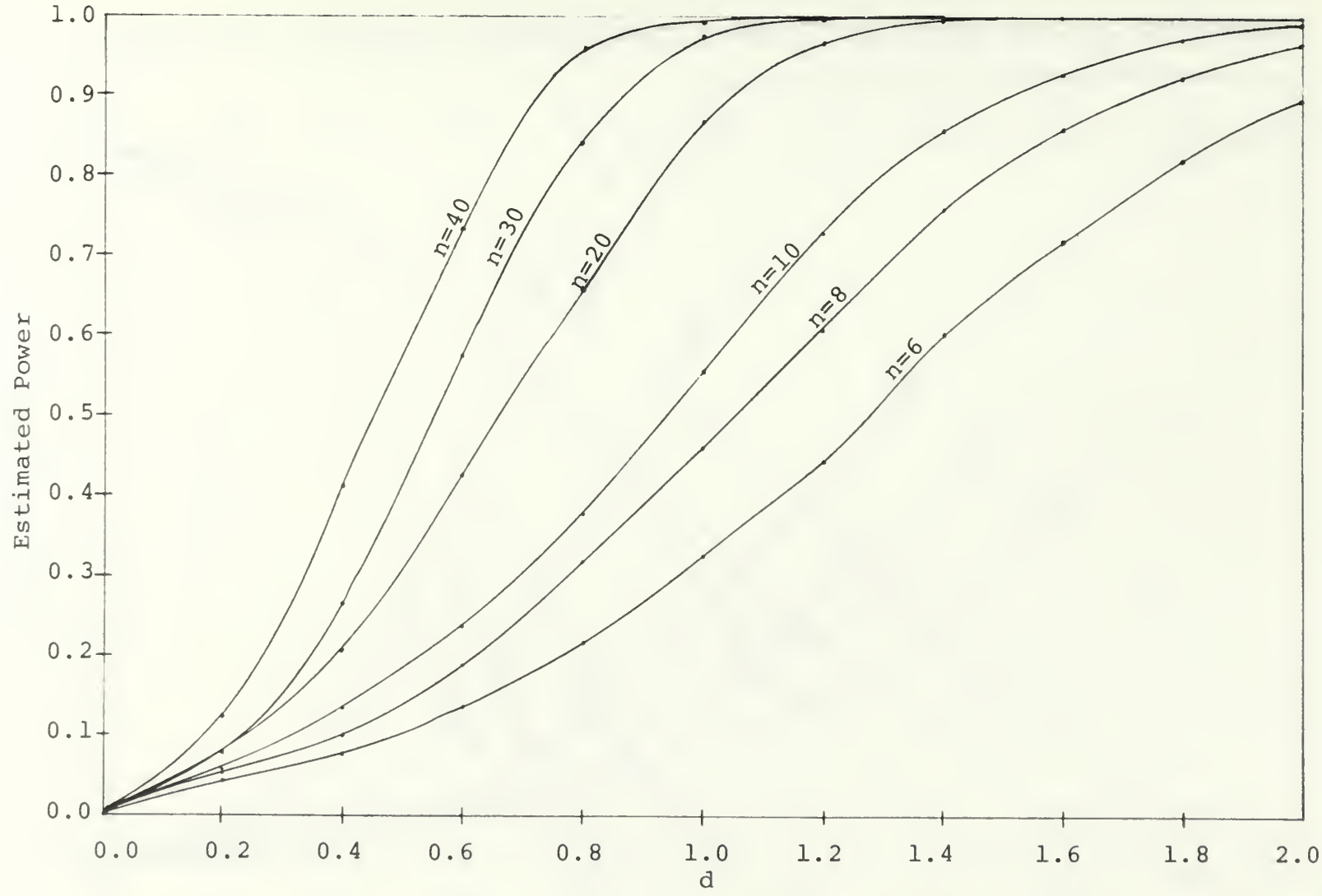

Figure 31. Estimated (unconditioned) power of 5% Duncan test of four means

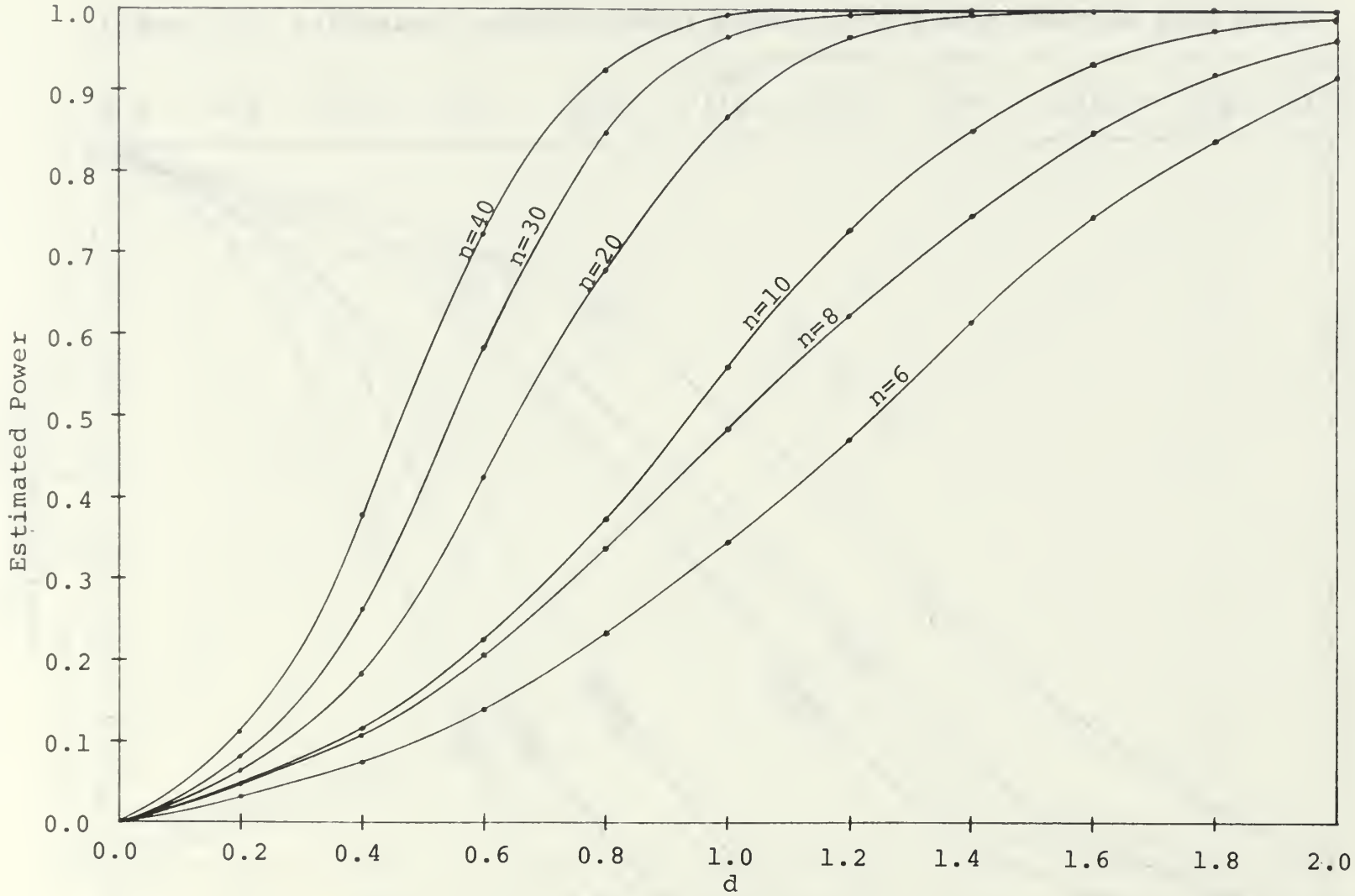

Figure 32. Estimated (unconditioned) power of 5% Duncan test of five means

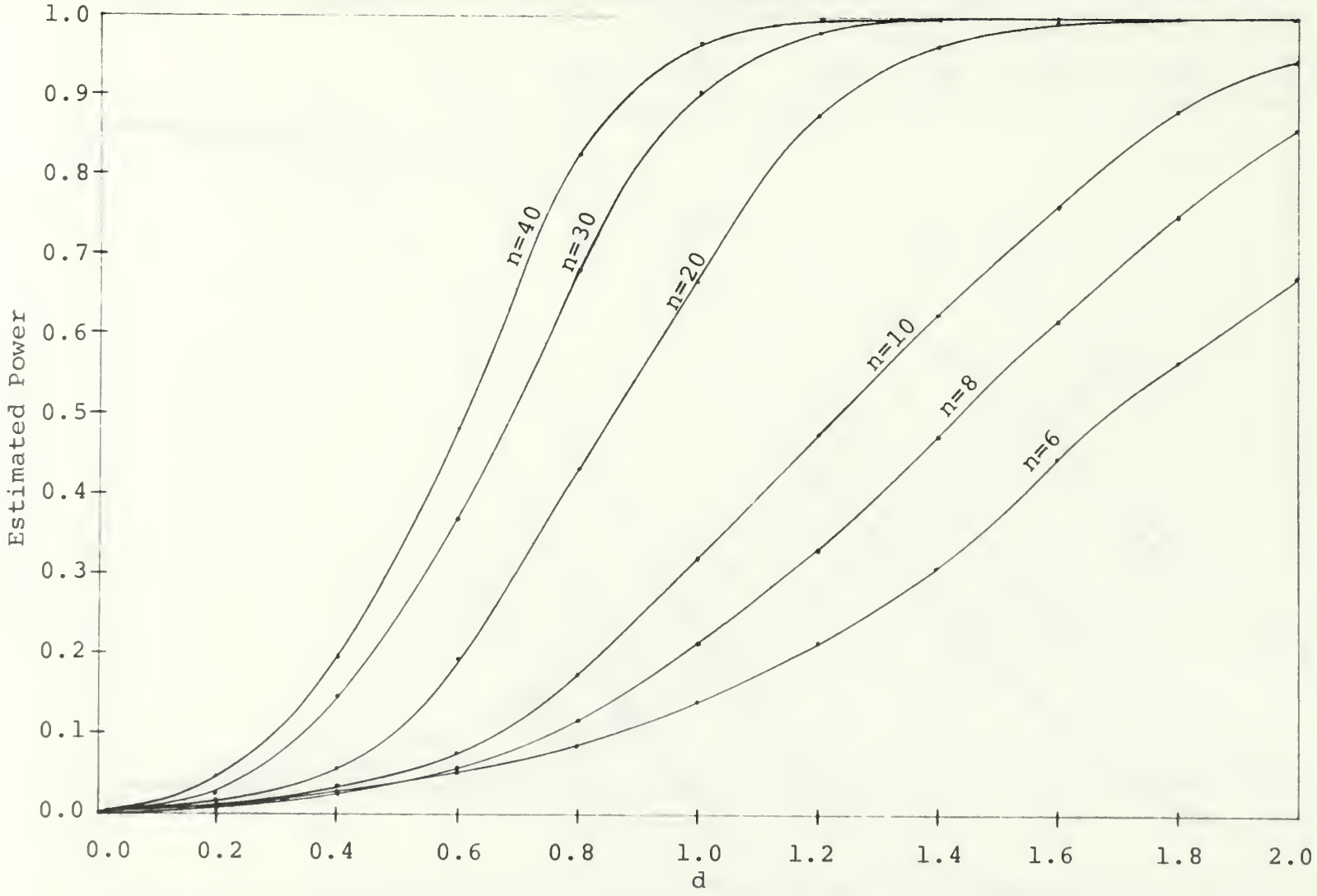

Figure 33. Estimated (unconditioned) power of 1% Duncan test of three means

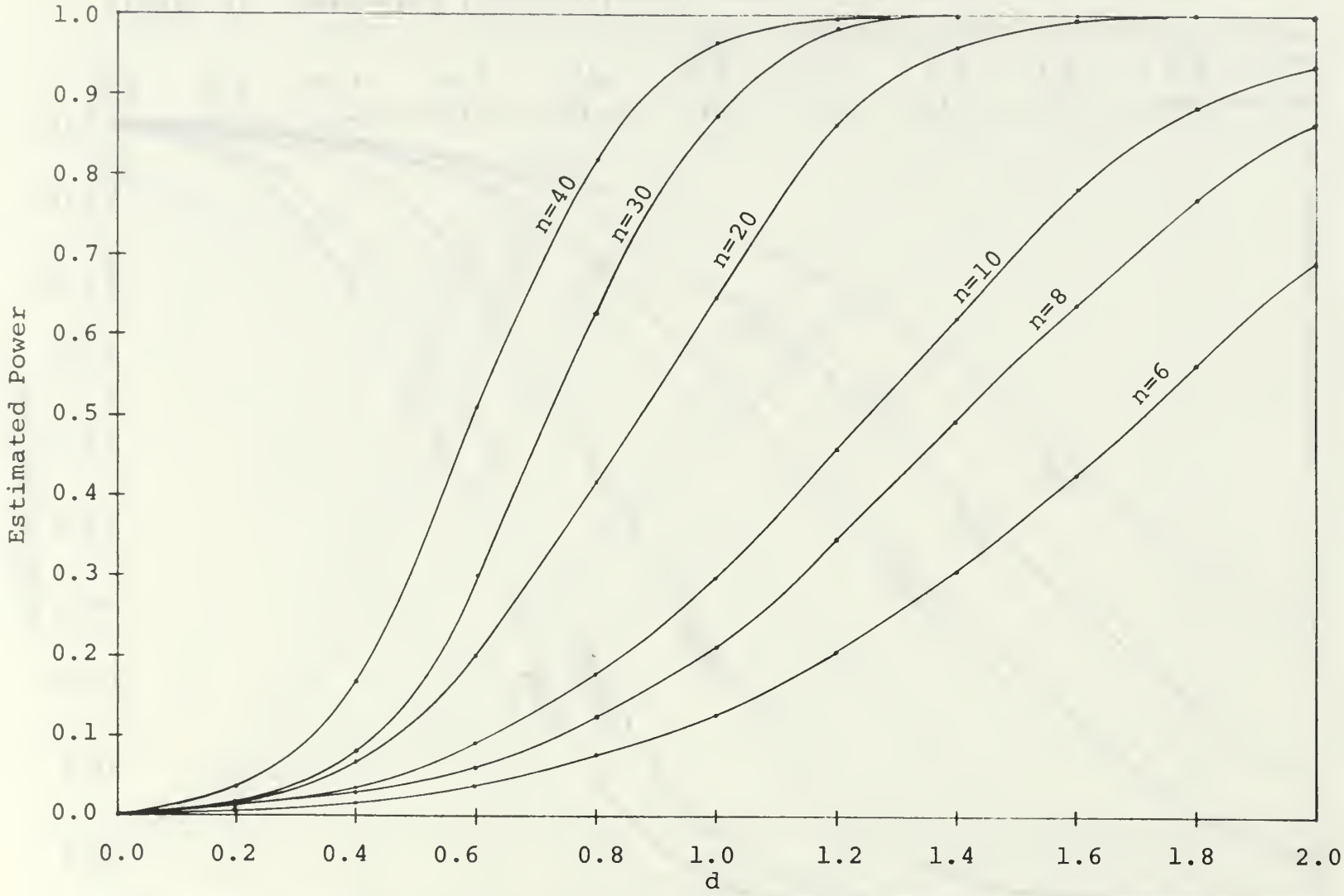

Figure 34. Estimated (unconditioned) power of 1% Duncan test of four means

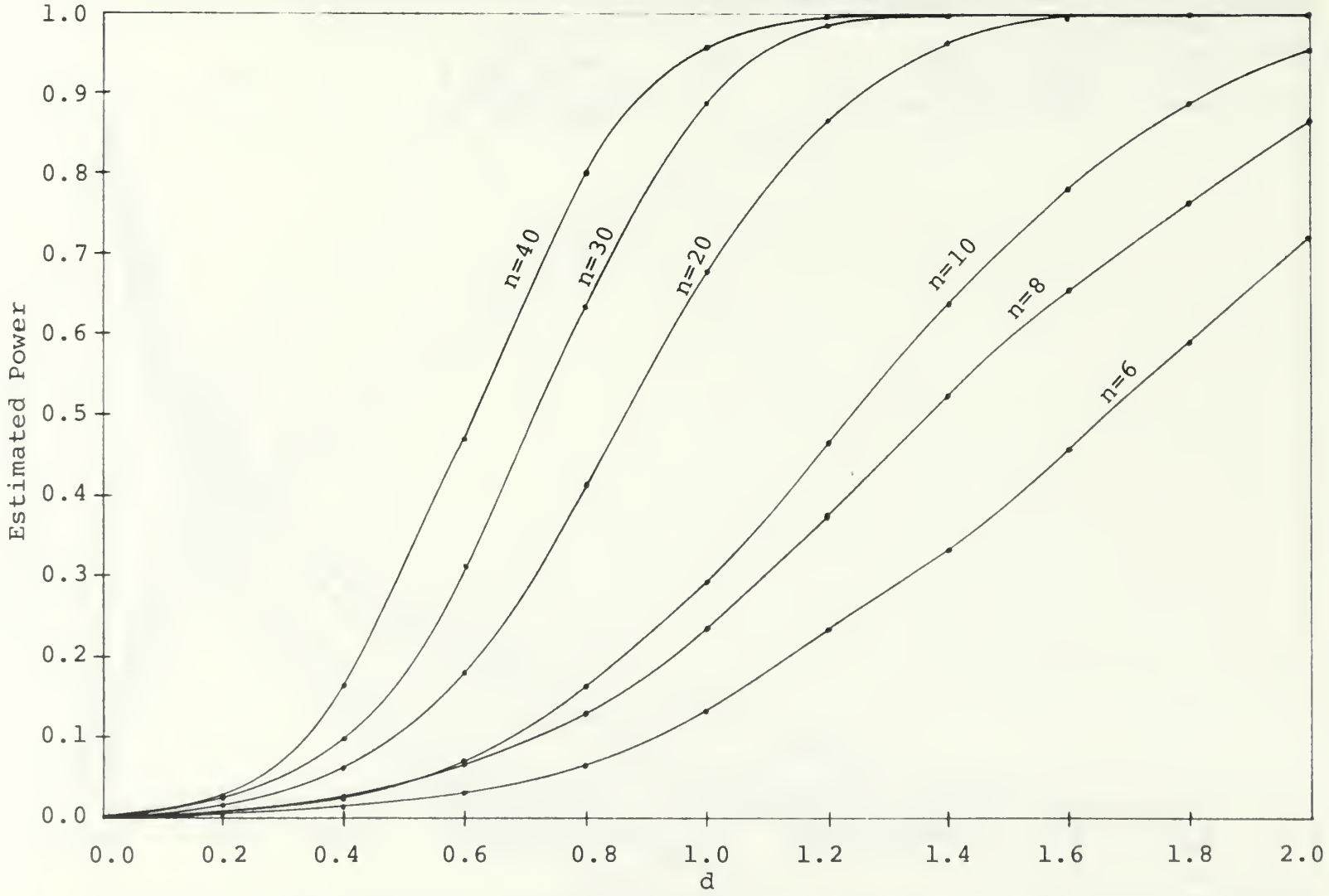

Figure 35. Estimated (unconditioned) power of 1% Duncan test of five means

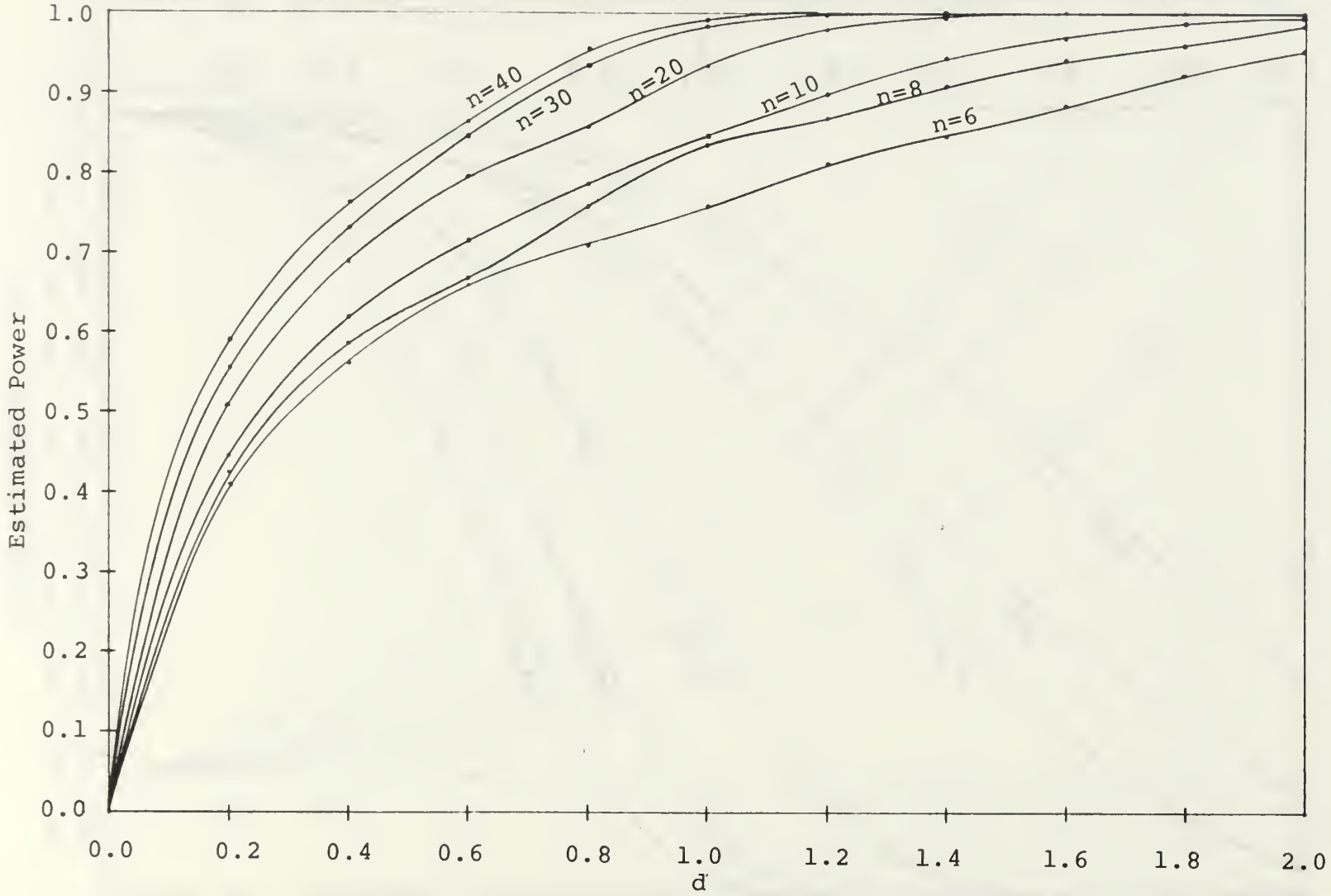

Figure 36. Estimated (conditioned) power of 5% Duncan test of three means

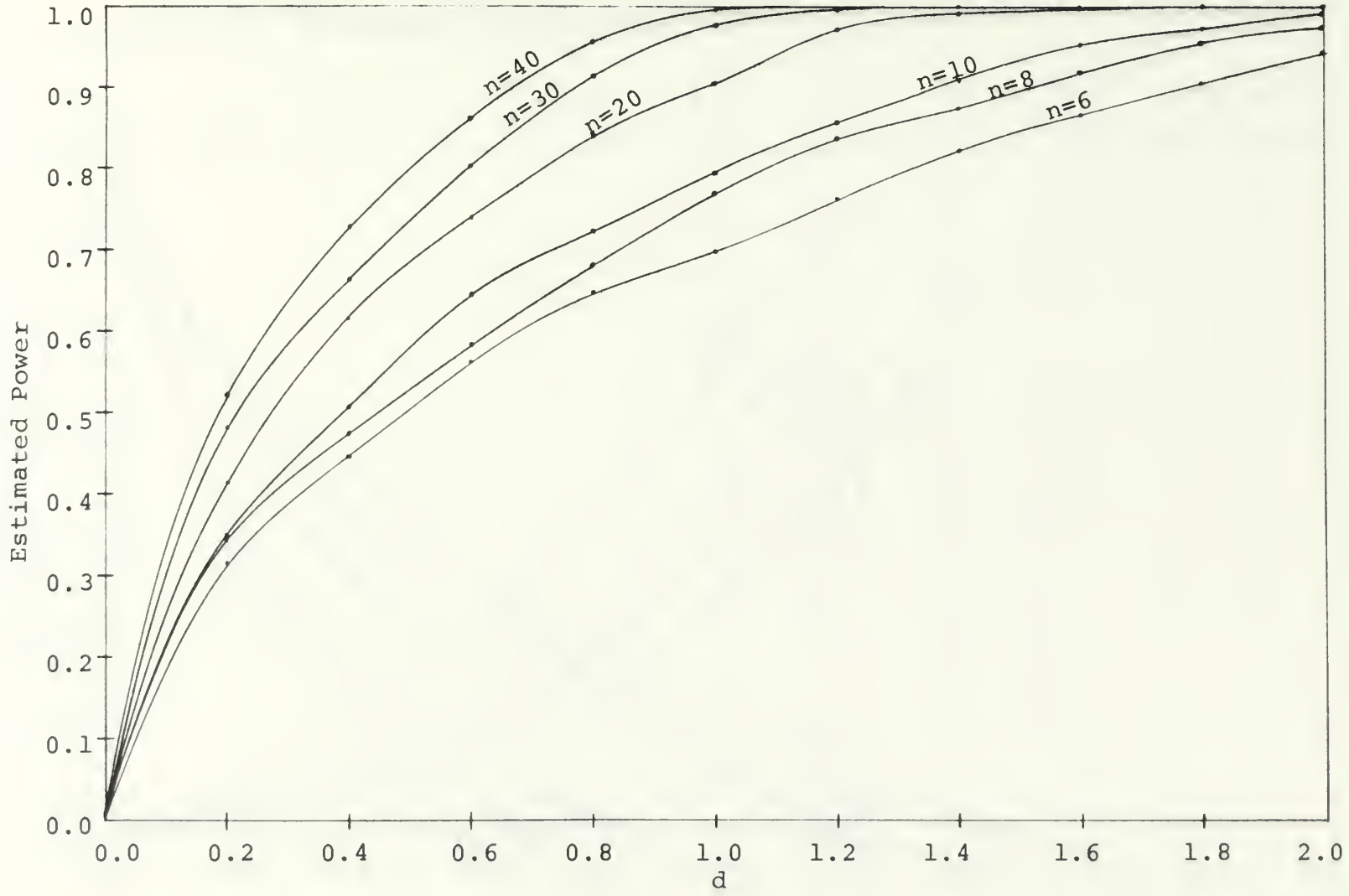

Figure 37. Estimated (conditioned) power of 5% Duncan test of four means

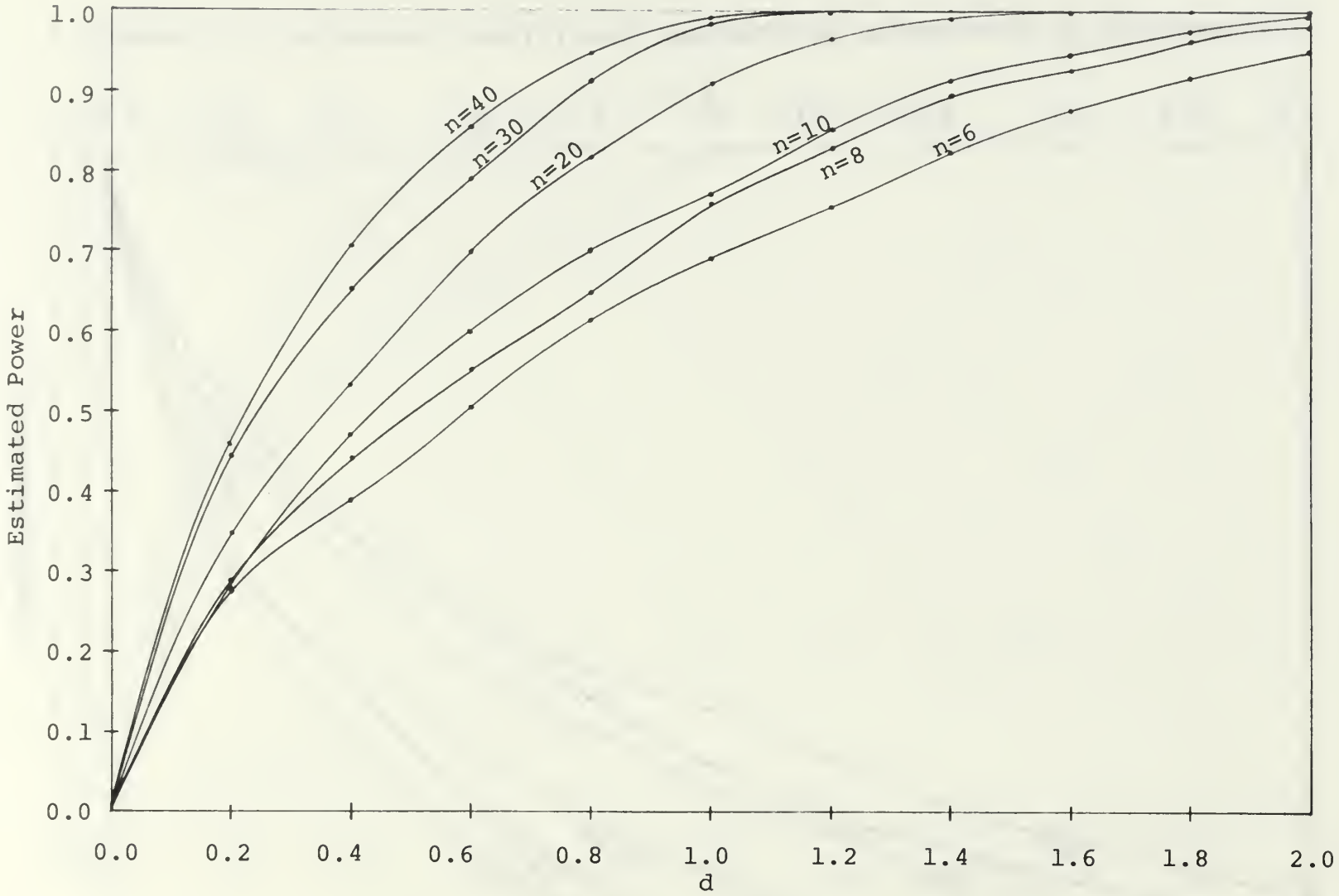

Figure 38. Estimated (conditioned) power of 5% Duncan test of five means

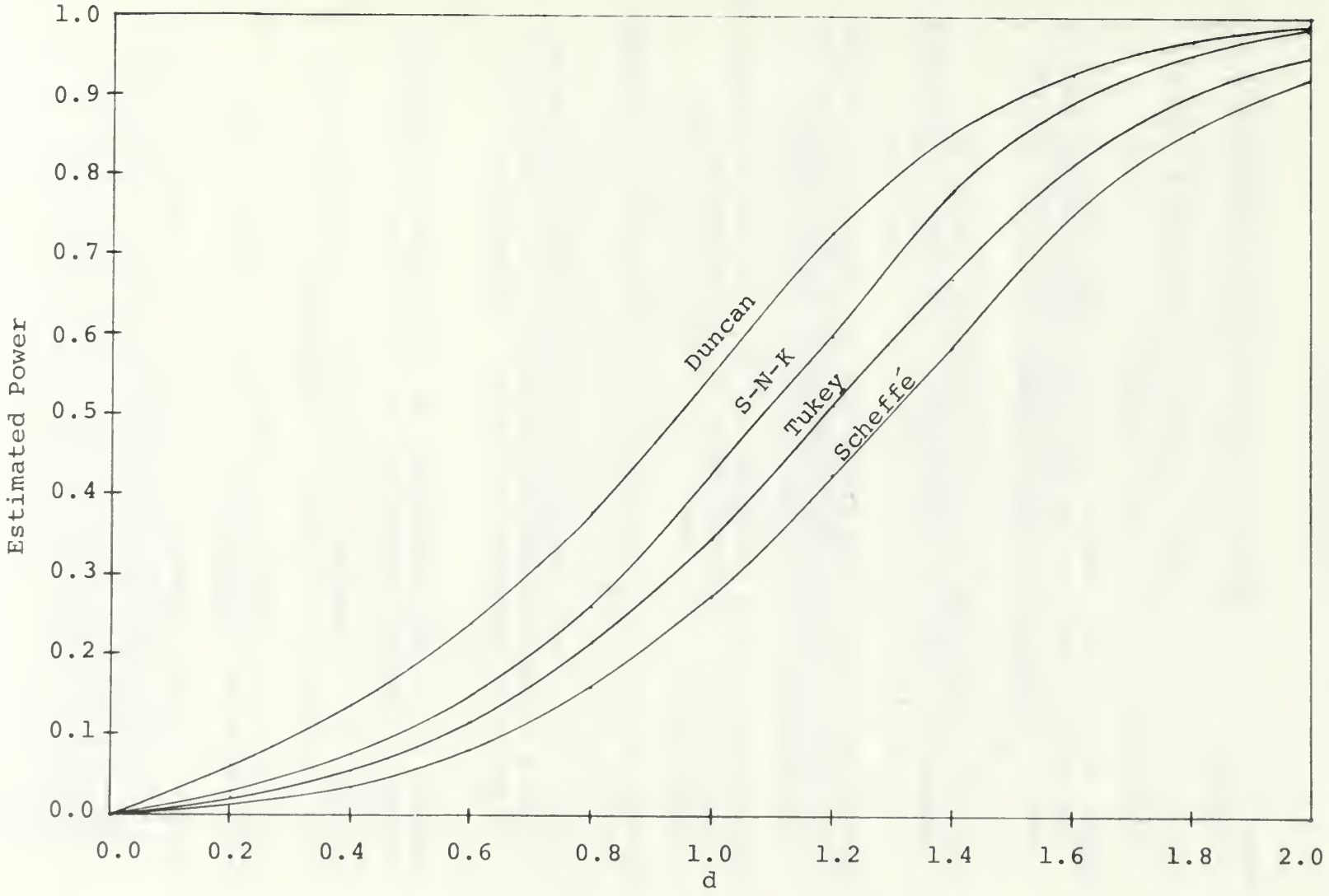

Figure 39. Estimated (unconditioned) power of 5% multiple comparison procedures for k = 4 and n = 10

BIBLIOGRAPHY

1. Bancroft, T. A., Topics in Intermediate Statistical Methods, v. 1, The Iowa State University Press, 1968.

2. Duncan, D. B., "Multiple Range and Multiple F Tests," Biometrics, v. 11, p. 1-42, 1955.

3. Green, B. F., Jr., Smith, J. F. K., and Klem, L., "Fmpirical Tests of an Additive Pandom Number Generator," Association for Computing Machinery Journal, v. 6, p. 527-537, 1959.

4. Keuls, M., "The Use of the Studentized Range in Connection with the Analysis of Variance," Fuphytica, v. 1, p. 112-122, 1952.

5. Marsaglia, G., "A Fast Procedure for Generating Normal Pandom Variables," Communications of the Association for Computing Machinery, p. 4-10, January 1964.

6. Merrington, M. and Thompson, C. M., "Tables of Percentage Points of the Inverted Beta (F) Distribution," Biometrika, v. 33, part I, p. 74-87, April 1943.

7. Miller, R. G., Jr., Simultaneous Statistical Inference, McGraw-Hill, 1966.

8. Newman, D., "The Distribution of the Range in Samples from a Normal Population in Terms of an Independent Fstimate of Standard Deviation," Biometrika, v. 31, p. 20-30, 1939.

9. Sarhan, A. E. and Greenberg, B. G., editors, Contributions to Order Statistics, p. 147, Wiley, 1962.

10. Scheffé, H., "A Method for Judging all Contrasts in the Analysis of Variance," Biometrika, v. 40, p.87-104.

11. Tukey, J. W., "Comparing Individual Means in the Analysis of Variance," Biometrics, v. 5, p. 99-114, 1949.

12. Wright Air Development Center Technical Report 58-484, v. II, The Probability Integrals of the Range and of the Studentized Range: Probability Integral and Percentage Points of the Studentized Range; Critical Values for Duncan's New Multiple Range Test, by H. L. Harter, D. S. Clemm, and F. H. Guthrie, October 1959.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	20
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	Director, Systems Analysis Division (OP 96) Office of the Chief of Naval Operations Department of the Navy Washington, D.C. 20350	1
4.	Assoc. Professor H. J. Larson, Code 55 La Department of Operations Analysis Naval Postgraduate School Monterey, California 93940	10
5.	LT Merlin Gene Bell, USN 671 Fast Drive Seymour, Indiana 47274	2
6.	Computer Facility, Code 0211 Naval Postgraduate School Monterey, California 93940	1

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Naval Postgraduate School Monterey, California 93940	Unclassified
	2b. GROUP

3. REPORT TITLE

A MONTE CARLO STUDY OF MULTIPLE COMPARISON TECHNIQUES

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Master's Thesis; October 1969

5. AUTHOR(S) (First name, middle initial, last name)

Merlin Gene Bell, Lieutenant, United States Navy

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
October 1969	87	12

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Naval Postgraduate School Monterey, California 93940

13. ABSTRACT

A study was conducted on the multiple comparison methods presented by Scheffe, Tukey, Student-Newman-Keuls, and Duncan under the experimental situation in which all populations were normal with equal variances and all means but one were equal. The characteristics of all four test procedures were compared for the case of multiple comparisons of pairs of means. These tests were conducted both with and without the prior performance of an analysis of variance. The Tukey and Scheffe procedures were compared in tests of linear combinations of three means. Estimates were made of the power of the tests and of Type I error rates under both the null and alternate hypotheses. Scheffe's method was found to be too conservative for pairwise comparisons of means, but it was to be preferred over Tukey's method for combinations of more than two means. Duncan's method was the most powerful test of pairwise comparisons, but it maintained little control over one kind of Type I error. The S-N-K procedure showed a good balance between power and control of Type I errors.

DD FORM 1 NOV 65 1473 (PAGE 1)
S/N 0101-807-6811

UNCLASSIFIED
Security Classification

A-31408

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
STATISTICS						
ANALYSIS OF VARIANCE						
MULTIPLE COMPARISONS						
MULTIPLE RANGE PROCEDURES						